DOCUMENT RESUME

ED 246 894 IR 011 226

AUTHOR	Mitton, Roger
TITLE	Practical Research in Distance Teaching: A Handbook for Developing Countries.
INSTITUTION	International Extension Coll., Cambridge (England).
REPORT NO	ISBN-0-903632-24-1
PUB DATE	82
NOTE	328p.
PUB TYPE	Guides - Non-Classroom Use (055)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Data Analysis; Data Collection; Developing Nations; *Distance Education; *Evaluation Methods; Foreign Countries; Observation; *Research Design; *Research Methodology; *Statistical Analysis; *Surveys
IDENTIFIERS	Lesotho

ABSTRACT

The linking of research and action, basic methods of social research, and the uses of research in distance teaching are examined, with examples drawn from the Lesotho Distance Teaching Centre and other distance teaching programs. Ways in which research can be useful and early stages of research design and planning are discussed. An explanation of basic research methods describes observation and discussion techniques and then details the steps involved in conducting social surveys, including sampling, creating questionnaires, survey fieldwork, data processing, and checking and reporting survey results. Experimental research is also addressed, and basic statistical concepts are introduced. Suggestions for applications of research in distance teaching settings relate to assessing instructional materials, pre-testing instructional materials, keeping regular records, and evaluation. Appendices contain additional information on statistical procedures, statistical theory, and costing, and an annotated list of 20 recommended readings. A four-page reference list, a list of potentially helpful organizations, and basic statistical tables are included. (LMM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED246894

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PRACTICAL RESEARCH IN DISTANCE TEACHING:

a handbook for developing countries

ROGER MITTON

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Jo Bradley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR011224

International Extension College 1982

International Extension College
18 Brooklands Avenue
Cambridge
CB2 2HN
England

© International Extension College 1982

ISBN 0 903632 24 1

Contents

page

Preface and acknowledgements iv
Introduction 1
How to use this book 5

Linking research to action

Chapter 1 How research can be useful 7
2 Before embarking on research 18

Basic research methods

3 Observation and discussion 27
Social surveys: introduction to chapters 4-9 35
4 Sampling 38
5 Questionnaires 48
6 Survey fieldwork 67
7 Processing the data 74
8 Checking survey results 98
9 Analysing and reporting survey results 109
10 Experiments 125
11 Statistical concepts 141

Applications of research in distance teaching

12 Assessing instructional materials 163
13 Pre-testing instructional materials 172
14 Keeping regular records (monitoring) 187
15 Evaluation: looking at projects as experiments 202
16 Evaluation: making value judgements 220

Linking action to research

17 After completing the research 233

Appendix 1 Statistical procedures 241
2 Statistical theory 272
3 Costing 283
4 Further reading 297
5 References 301
6 Addresses of organisations 305
7 Statistical tables 307

Index 313

PREFACE AND ACKNOWLEDGEMENTS

On my return to England in 1977, after spending three years in Lesotho, Africa, as deputy director of the Lesotho Distance Teaching Centre, I was commissioned by the International Extension College to write a book about research in distance teaching. The original idea was that I would write a short book of about 60 pages which IEC would publish in their series of broadsheets. As I got into it, however, it became clear that, if I was going to explain in detail how to do research, the book would have to be much longer. Since I was not able to work full-time on the book, progress was slow. The first draft was eventually completed in 1981 and the final draft in 1982.

Of the people who helped me, two deserve special mention - Alan Etherington, formerly evaluator at the Botswana Extension College, and Rick Powell, formerly evaluator at the Lesotho Distance Teaching Centre. I discussed the book with them at the beginning and received further advice from them on several occasions while writing it. I also received help from Philip Baker and Conrad Halloran with the section on costing, and from Michael Willmott with the chapters on statistics.

Every year a number of people from distance-teaching organisations in developing countries attend a course in London run by the International Extension College. Several of the students on these courses read preliminary versions of some chapters and gave me their comments. Various staff members of IEC also gave assistance, especially Jo Bradley, who read the manuscript, made some editorial improvements and supervised the publication. IEC also compiled the list of organisations in Appendix 6. The typing was done by Sandra Last and Maureen Stirling.

In addition to those already mentioned, the following gave me their comments on the first draft: Barbara Baddoo, David Crowley, Geoff Dench, Tony Dodds, Malcolm Ford, Jim Hoxeng, Janet Jenkins, Roger Lewis, Bob Mackenzie, Motlatsi Morolong, Paud Murphy, Hilary Perraton, Mary Thorpe, Wyn Tucker and Peter Willmott.

Janet Jenkins, Arthur Lucas and Peter Burclaff helped with the illustrations. Those that come from publications of the Lesotho Distance Teaching Centre are reproduced by permission of the director, and the statistical tables on pages 307-312 by permission of the original publishers.

The opinions expressed in this book are my own. They are not necessarily shared by the International Extension College, the Lesotho Distance Teaching Centre or the people who have helped me with the book.

Introduction

Distance teaching is any kind of teaching in which, for the most part, the teacher communicates with the learners through some medium such as print or broadcasting, though he may occasionally talk to them face-to-face. Correspondence courses, educational radio or television broadcasts, instructional booklets, pamphlets or films are all forms of distance teaching.

Research can help at every stage of a distance-teaching project. In the early part, research can help to decide what to teach, to whom, and how to go about it. When a project gets under way, research can help to develop the materials and to monitor the running of the project. After the project has been running for some time, or after it has finished, research can help in evaluating its effect.

By 'practical research' I mean research which is undertaken to help a distance-teaching organisation do a better job. Someone is supposed to do something in the light of the results - to change a policy, to redraft some material, to launch or cancel a project, or whatever. And the researcher carries out his research with that in mind from the outset. This is a bit different from doing research with the main purpose of publishing an article in a learned journal or fulfilling the requirements for a degree, when the researcher does not necessarily intend that anyone should take action on the basis of his findings.

This book is about doing practical research. It is not a summary of research findings on distance teaching, nor is it a digest of the literature on educational research and evaluation. It is the advice that I would give if someone came to me and asked how to go about doing research in distance teaching.

My primary audience is people doing distance teaching in developing countries. They might be working in a specialist distance-teaching organisation, such as a correspondence college or a media production unit. They might be providing a distance-teaching element to supplement face-to-face teaching, such as radio programmes for schools or leaflets to accompany the lectures given by agricultural extension agents. Or they might be doing some distance teaching as part of an activity that is not primarily educational at all; for example, an organisation distributing food might want to publish a pamphlet on nutrition. This book will also be useful, I hope, to distance-teaching organisations in the richer countries. In fact, much of it is relevant to educational research and evaluation in general, not just in distance teaching. But it is mainly distance teaching in the developing countries that I have had in mind.

My own experience of research began at university, when I conducted an attitude survey among the students as part of my work for a degree in psychology and philosophy. My first job on leaving university was to help write the story of a community development project that had been conducted in the 1960s in a part of London. Then I worked for two organisations in London which specialise in social research, mainly working on social surveys. My main qualification for writing this book, however, is the three years I spent in Lesotho, Africa, from 1974 to 1977. I have tried to make this book more widely relevant, by reading books, talking to people and showing drafts of the book to people with experience of other countries. But, inevitably, I have relied mainly on my own experience, so I should say a little about Lesotho and my work there.

Lesotho is a small country in southern Africa. It is roughly rectangular, about 200 km long and 130 km wide. On the map, it looks like an island, since it is completely surrounded by the Republic of South Africa. In fact, it is more like a mountain stronghold than an island, enclosing within its borders the highest mountain range in Africa south of Kilimanjaro. The people, who numbered about 1.25 million at the 1976 census, are called Basotho and their language is Sesotho. The country was the British Protectorate of Basutoland for almost a hundred years, gaining its independence in 1966, so English is the language of the secondary schools and the civil service. Lesotho has maintained its political independence from South Africa ever since its origins early in the last century. Economically, however, it has to be dependent on its rich neighbour, having few agricultural or mineral resources of its own. Most of the men work in the mines in South Africa, doing a 12-month or 18-month contract, then coming home for a while, then going back to do another contract. The Christian churches are strong in Lesotho, the earliest missionaries having arrived not long after the nation was founded, and there are many missions and schools dotted round the country. The capital city is Maseru, population about 45 000, and there are a few other small towns, but the great majority of the people live in small villages of round, thatched houses.

I arrived in Lesotho in May 1974 to join the Lesotho Distance Teaching Centre (LDTC). The Centre had been established a few months earlier, but it was still very small and short of money. My arrival brought the staff total to four. The Centre had been set up with the broad aim of exploring ways in which distance-teaching methods might be used to make education of all kinds more widely available in Lesotho. Despite my research background, I was not appointed as a researcher. I was deputy director, and my job was to help to get the Centre established and, in particular, to organise the Centre's work in rural education. However, in trying to decide what we should do and how we should do it, I was inclined to think of research as a way of approaching the problems. So while I spent most of my time helping to attract funds, recruit staff, set up the organisation, produce our first materials and so on, I also conducted a few pieces of research. Later, a full-time researcher was appointed to take over this work, and eventually a small research division was established as a permanent part of the Centre, with a staff of three out of the Centre's total staff of about fifty.

The Centre's work was varied. We offered courses to people studying at home for the Junior Certificate and the General Certificate of Education 'Ordinary level' (the exams that schoolchildren take after three years and five years of secondary education). The teaching for these was mainly by correspondence courses, supplemented by radio programmes and occasional sessions of face-to-face tuition. We published booklets for rural people, on topics such as vegetable gardening and child care. We produced pamphlets, posters, radio programmes and radio spots for other agencies, such as the Lesotho Family Planning Association and various rural development projects. We designed teaching aids for the fieldworkers of other agencies and ran short courses for the fieldworkers in how to use them. We collaborated with the National Teacher Training College in a teacher upgrading project. And we ran experimental courses in basic literacy and numeracy for young people, using workbooks and games. We did not use television, because Lesotho did not have a television service, and we never made films, though we did use videotape in training fieldworkers.

In this book I have drawn many examples from the work of the LDTC and I have occasionally used examples from elsewhere, particularly from my colleagues in Botswana. Most of the examples are real-life ones, but I have occasionally removed or changed certain details, sometimes to make the example clearer and at other times to avoid giving offence.

The first two chapters and the last are about linking research and action. The rest of the book is about doing research. Chapters 3 to 11 describe the basic methods of social research; it is not only researchers in distance teaching who make use of these methods - they are the basic tools of social research in general - but I use examples from distance teaching to illustrate them. The remaining chapters - 12 to 16 - look at some of the tasks that research can perform in distance teaching, such as pre-testing instructional materials.

A glance at the contents page will show that I devote a lot of space to social surveys. I have to admit that this partly reflects the bias of my experience - I have done many social surveys and I have a lot to say about them. But there is a better reason: the social survey is particularly well suited to the kind of problems that arise in distance teaching in a developing country. In a health education campaign, for example, the typical audience is a large number of people scattered over a wide area, and the researcher's task is to find out about them. The social survey is a research method devised for precisely this kind of task. A survey could be conducted before the campaign, to guide policy, help with materials design or provide a baseline for evaluation; further surveys could be conducted during the campaign to monitor progress, or afterwards to evaluate its impact. In the first few years of the LDTC we found the survey to be the most appropriate method for many of the research problems we faced; the space I devote to it in this book reflects the extensive use we made of it there.

Most of the research techniques described in this book have their origins in European or American social science; I am recommending that they be applied in developing countries and this raises a question. One cannot simply assume that the practices of one's own society can be, or should be, exported, and this presumably includes social research techniques. And there are several writers who - if their writings are to be taken literally - think that what I am recommending in this book is all wrong. Some feel that social research is of such doubtful validity that it is not worth doing at all. Some feel that studying people - treating people as the objects of research - is somehow unethical. Others argue that, though these techniques may be valid and acceptable in the societies for which they were designed, they should not be applied in other, quite different societies. Appendix 4 gives references to some articles where these views are put forward.

I have decided not to state at length my arguments against these writers - this book is already much longer than I wanted it to be. Suffice it to say that, while I accept many of their criticisms of social research in general and in particular of the attempt to apply these research techniques in developing countries, I find their arguments too theoretical and too negative. If you have a job to do, such as finding out what effect a health education campaign has had, you need some idea of how to go about it. There are techniques that you can use. They are not perfect and they will need to be adapted for your particular problem. But having these techniques is better than having none at all.

Whatever the balance of the theoretical arguments, the decisive test of these research techniques is how useful they are in practice. I give many examples in this book of these techniques being applied in developing countries. It seems to me that they work, that there is generally nothing unethical about them and that they can produce useful results. In short they can be of value provided that they are applied with sensitivity and common sense, and the purpose of this book is to show how this can be done.

How to use this book

I recommend to both researchers and non-researchers that you should read Chapters 1, 2 and 17, even if you don't look at anything else. These are about integrating research with the other activities of a distance-teaching organisation so that the research really helps to guide the organisation's work.

Of the rest of the book, the parts that are more likely to interest non-researchers are Chapters 12-16. Anyone involved in materials production - writers, editors, artists, radio producers and so on - should read Chapters 12 and 13. People administering correspondence study, such as student advisers, should read Chapter 14. Chapters 15 and 16 will be of interest to most people, especially project directors.

Researchers can use it as a textbook or a handbook. If you just want to learn about research, you can read it straight through - the material is organised so that you take things in a logical order. A few parts are more specialised and you could skip these if you felt they weren't relevant to you - they are printed in a smaller typeface, like this:

Heads of households in rural Lesotho (an example of sampling)

Different countries will present different sampling problems, of course, but I will illustrate the points I have made by describing how we might take a sample of rural heads of households for a survey in Lesotho. (I will simplify the details, but only slightly.) If you don't want this amount of detail, skip this section.

Most readers, I imagine, will use it as a handbook to dip into for advice on particular topics. If you know precisely what you want - the procedure for doing the Mann-Whitney U test, say - use the index. If you want advice on a general subject rather than a precise topic - writing a questionnaire, say - you can try the index or you can select the most promising-looking chapter from the contents list. Each chapter begins with a preview - a section in this typeface:

Observation is a good research method if you want to give yourself a picture of people, places or events. To get the most out of an observation visit, make notes in advance and write a report as soon as possible afterwards. Participant observation.

This lists the chapter's sub-headings and gives a very brief indication of what's in each section. If you come across a technical term that is not explained, it's probably explained somewhere else in the book, so look it up in the index.

I have tried to include everything you need to know in order to do an adequate piece of research, but I have not attempted to include everything you could possibly want to know. Appendix 4 suggests some books and articles you can turn to for further advice and several of them have long bibliographies for yet further reading. If one of the real-life examples that I mention catches your interest, you can find the reference to the report on it, if there is one, in Appendix 5. Copies of many of the LDTC reports are available from the LDTC - details at the end of that Appendix.

1 How research can be useful

This chapter describes several ways in which research can be useful in a distance-teaching organisation, and illustrates them with examples from the LDTC (Lesotho Distance Teaching Centre).

Policy guidance Before designing a distance-teaching project you can find out about the potential audience, about the media you might use and about other organisations doing similar work.

Materials design By testing alternative versions of the same material you can provide guidelines to designers on the style or technique they should employ.

KAP (Knowledge, Attitudes, Practices) Before trying to educate people about something it is a good idea to find out what they already know about it, how they feel about it, and what they do about it.

Pre-testing Before publishing written material or broadcasting radio programmes you should test the material on a few people to see if it conveys what it is supposed to convey.

Monitoring Collecting information at regular intervals about an ongoing project can warn you if things are going wrong and can also provide a baseline against which to judge the impact of innovations.

Evaluation Research helps you to see to what extent your distance teaching is having the desired effect and more generally to assess its value.

Appropriate research I am not proposing massive research projects. I am recommending small-scale pieces of research closely tied to the work of the distance-teaching organisation.

Policy guidance

Policy-makers need facts. In the early days of a distance-teaching organisation, or even before the organisation is established, policy-makers have to take big decisions about the type of work the organisation is going to do and the way it is going to do it. Thereafter, they are in a similar position every time the organisation takes on a new project. They can use research to provide themselves with the facts they need.

In the absence of these facts, the decisions are likely to be over-influenced by other things. The director of a new organisation will be strongly influenced by his previous experience; if he has had great success elsewhere in using videotape to stimulate village discussion of local problems, he is quite likely to think, when he takes up his new post, that what is needed is village discussion stimulated by videotape. Or a director can be over-impressed by ideas he has picked up on courses; if he has seen puppet shows used to great effect by other organisations - perhaps in other countries - he may want his own organisation to try them.

This is natural and, to an extent, a good thing. A new director is expected to bring the benefits of his experience and to pick up ideas on courses. But these influences ought to be counter-balanced by facts about the country he's working in. To give a simple example, it would be foolish for a new director to devote a large part of the organisation's resources to television, because of his personal enthusiasm for that medium, if there were very few television sets in the country.*

One cannot say in general what facts an organisation should find out. It all depends on the particular policy questions that the organisation faces. But some examples from the Lesotho Distance Teaching Centre might show the kind of facts that research can provide.

Early in the rural education work we had to decide how much use to make of radio and printed materials. If we transmitted a radio programme, how many people would be likely to hear it? If we distributed a leaflet, how many people would be likely to read it? We conducted an interview survey of 250 rural adults and found, among other things, that about half the people could read a simple text, whereas the proportion who would be likely to hear a radio programme was rather small, probably less than a fifth.**

Before committing ourselves to offering correspondence courses, we carried out a small test of the postal system. We sent letters to the headmasters of several schools around the country enclosing a stamped addressed postcard. We asked them simply to post the cards straight back to us. We discovered that the postal system worked very well; most letters were delivered in two or three days, although letters to the mountain districts could take up to two weeks.

Since other organisations were already engaged in offering practical instruction to rural people, we conducted an informal survey of them to find out what they were doing and how we could fit in with them. We put similar questions to all of them (there were about thirty organisations) about their aims and methods, the number of staff they had, and so on. One thing we discovered was that almost all of them relied exclusively on the lecture method of instruction and made little use of support materials such as pamphlets or visual aids. This was because their stock of materials had generally been produced for other countries, not specifically for Lesotho, and they did not have the equipment or expertise to produce their own. Many of them felt the need of an agency which would design and produce materials locally for use in nonformal education. As a result of this, LDTC tried to fill this role. We provided this service over the next few years, to about twenty organisations.

---

* Throughout the book I will refer to characters such as 'the director', 'the researcher', 'the editor', 'the artist' and so on by using the pronoun 'he'. I am using it as an abbreviation of 'he or she'. I do not mean to suggest that such positions are held, or ought to be held, solely by men.

** Reports on this survey and on other research mentioned in this book are available from LDTC or from other organisations. References are given in Appendix 5.

Because we had discovered that a high proportion of rural adults, especially housewives, could read, and also that there was very little printed material on practical topics in the Sesotho language, we decided to experiment with a range of booklets, simply written, cheap and practical. We wondered what topics to choose. The obvious thing to do was to ask the housewives what topics they were interested in. We had already agreed to produce 10 000 cookery booklets, in response to a request from another agency, so we included stamped addressed postcards in these, asking the readers to tell us which topics they would like us to choose for future booklets. Although only a small proportion of the readers sent us the cards, the order of preference was clear. Not surprisingly, child care was the most popular topic, followed by crochet and vegetable gardening.

One of the questions that arose early in the preparation of courses for examination candidates was how much help the students would need in addition to the correspondence courses. (If, for example, many of them found it too noisy to study at home, perhaps LDTC could arrange for them to make use of school classrooms out of school hours.) To answer this question, we arranged for the Examinations Council to include a questionnaire from us along with the official form that private candidates had to fill in when applying to take the examination. The responses to this questionnaire gave us a good picture of the circumstances in which these candidates were trying to study at home.

As I mentioned at the beginning of the chapter, one may need this fact-finding research even when the organisation is established, if one embarks on a new project. The Lesotho Family Planning Association was one of the organisations that took up our offer of help with the design and production of materials. They asked us to produce support materials, and training in how to use them, for their twenty fieldworkers. This was a new project for us. We needed to know how the fieldworkers actually did their work, what problems they had and what sort of materials they thought they needed. We visited about ten of the fieldworkers and found that the main part of their work consisted in giving fairly formal lectures to village meetings. We also attended one of these meetings. This initial research was very useful when we designed the materials - a flipchart to accompany the lectures, and pamphlets reinforcing the same points. For example, we found out what aspects of family planning were felt to be unacceptable for public discussion at a village meeting.

These examples illustrate the diversity of facts that one might need in setting up an organisation or designing a project, and also some of the ways in which research can provide these facts. These particular examples relate to LDTC's policy questions; other organisations will have other problems and will need different facts. But the general questions that underlie these examples are likely to apply to other organisations:

> We needed to find out about the potential 'students' (private candidates, farmers, housewives, or whoever), about their interests, their abilities, their problems.
>
> We needed to find out about the media that we might use (print, radio, postal services).
>
> We needed to find out about other agencies who were doing the same kind of work.

Materials design

Writers, editors, illustrators and scriptwriters often feel the need for guidelines, especially in the early stages of drafting or designing materials. A course writer might want to know whether to include self-check exercises. An illustrator might be unsure whether to use photographs or line drawings. A radio scriptwriter might wonder whether to present his material as a straight lecture or in drama form. Research can sometimes help.

At LDTC, we tested photographs against line-drawings by producing a set of pictures in both styles, e.g. a photograph of a bus and a line drawing of the same bus. We showed these to a large number of rural people, in individual interviews, and asked them to say, in each case, what it was a picture of. We found that people did not do consistently better with any particular style. The photograph was better for some items, where shading or texture helped people to recognise the object, while line drawings were better for other items, where some small detail was crucial for correct interpretation.

We also conducted a test to assess the effectiveness of various additions to a correspondence lesson. We wrote a basic lesson, which simply presented the information in straight text. Then we produced different versions of this basic lesson, each one incorporating a learning aid - a lively introduction to capture the attention, a detailed statement of the objectives of the lesson, questions with immediate answers, more general questions, or a summary. We gave each version to a group of schoolchildren and gave all the children the same test at the end. We found that none of the learning aids made much difference, though there was some indication that the questions with immediate answers had a slight effect on the students. There had been some debate among the writers about whether to adopt this or that particular learning aid, and the effect of this unexciting research result was to take some of the heat out of the argument.

Another piece of work of this kind concerned some radio programmes we were broadcasting, which were intended to help people to pass an examination in book-keeping. The scriptwriters wondered how much they should encourage the listeners to take notes. To test this, we produced two different versions of the same programme, one which suggested in a general way that the listeners might take notes, and the other which explicitly instructed the listeners to take notes, told them which words to write down and gave them time for the writing; the second version was built around the note-taking, whereas the first was not. We played the first version to one class of schoolchildren and the other version to another class, using cassette recorders. We had given a test to each class before the programme and we gave them the same test after it. We found that the class who had heard the second version had taken better notes and that they did better on those test items that required them to remember certain terms. However, they did not do better on those test items that required them to use those concepts. It seemed that they had learnt the terms better, but had not understood the concepts better.

Often, it is not necessary to conduct one's own research to answer general questions about materials design. A lot of research on these questions has been published and you might find that someone else has already discovered the answer to your problem. There are various documentation centres (listed in Appendix 6) who might be able to refer you to the appropriate report.

At the same time, you must think carefully about whether research conducted in another country, and perhaps for a different purpose from yours, is applicable to your particular problem. For example, one researcher reported that people found it easier to interpret a blocked-out photograph (a photograph of an object with the background removed) than a line drawing and, therefore, used blocked-out photographs in his posters and flipcharts. Some people have concluded from this that they should always use blocked-out photographs and never use line drawings. This is silly. Line drawings might be adequate, even preferable, for certain purposes. Research reports from elsewhere can be useful in giving you ideas or warning you about problems you had not thought of, but it is a mistake to adopt their recommendations uncritically.

Knowledge, Attitudes, Practices

If you intend to give people practical advice, for example on poultry keeping, vegetable gardening or family planning, it is useful and sometimes essential to find out first what the people already know about this topic, how they feel about it and what they do about it. This is sometimes called a KAP study - Knowledge, Attitudes, Practices.

Two examples will show the value of this. We conducted a KAP survey of poultry keeping and found that most poultry owners kept two or three birds and said they might be interested in buying two or three more. To be of use to them, therefore, the advice should be about keeping a small number of birds. This was a useful finding since the technical advice we had obtained from poultry experts was about how to keep fifty or more. Much of this advice would clearly be inappropriate.

We also conducted a KAP survey of family planning. We found that most people thought that if a woman had sexual intercourse while she was breastfeeding, her milk would go bad and poison the child. This helped to explain why many people were hostile towards the idea of family planning. Family planning educators were recommending the use of contraception as a method of child-spacing. In the eyes of many people, this was like recommending that a breastfeeding mother should poison her child. We concluded that if family planning educators recommended contraception for child-spacing, they would also have to present the facts about the effects on breast milk. The two topics were closely related in people's minds; to talk about one and not the other was just inviting trouble.

Pre-testing

Instructional materials should always be pre-tested to make sure that the intended audience can understand them in the way they are supposed to do. I am using the word 'pre-test' here to mean that the materials are tested before they are put to use in a full-scale educational programme. You can ask some people to read a draft pamphlet or look at a draft illustration and then ask them questions about it. You can ask some people to work through a draft correspondence lesson and then give them a test on it. You can ask some people to listen to a tape of a proposed radio spot or radio programme and then discuss it with them. On the basis of the pre-test results, the writer or artist can alter the material before it is printed in large quantities; the radio producer can modify the programme before it is broadcast.

Writers and artists are inclined to think that their work is perfectly clear, that people could not possibly misunderstand it and that pre-testing is unnecessary. But pre-testing almost always suggests some ways in which the material can be improved and sometimes indicates serious faults in the material.

Sometimes people have difficulty with a complicated sentence or with an obscure word or expression. This is particularly likely to happen in a country which uses two languages - the local language and an international language such as English - and when the material is either in the second language or translated from the second language. When pre-testing an agricultural leaflet, for example, we found that some semi-technical terms for various types of soil made no sense when translated into the Sesotho language for the local farmers.

Sometimes the reproduction quality is not clear enough. A badly printed photograph can cause difficulty, or a poor sound recording. On a radio spot about nutrition, for example, we had some sound effects of someone cooking. Though these noises were reasonably clear if you already knew what they were, the pre-test showed that they were not at all clear to other people.

Sometimes people misinterpret material in a way that is reasonable yet quite unexpected. A draft pamphlet on family planning, for example, contained a drawing of a husband and wife talking to a nurse at a clinic. In the pre-test, many people thought that the couple were not married because the artist had put them on opposite sides of the table. Others did not realise that the third person was a nurse, so they thought the people were at home, or in school, or in a police station. Once we had discovered these problems, we could see how easy it was for people to make these mistakes and how the picture needed changing, but none of us had seen these problems before the pre-test.

A simple pre-test can sometimes show up a small but crucial flaw in the material. I had drafted a booklet which explained how to crochet. It contained sentences such as, 'Hold the crochet hook between thumb and first finger.' For the first pre-test, I just gave the draft to a staff member, with a crochet hook and a ball of wool, and asked him to follow the instructions. He had great difficulty even in just holding the hook and the wool correctly. I asked him why he was not following the instructions and he protested that he was. It turned out that, in Sesotho, the fingers are numbered differently; by 'first' finger, I had meant the index finger, whereas he called his little finger the 'first' finger.

If materials are not pre-tested, there is a serious risk that the people on the receiving end will receive either the wrong message or no message at all. I strongly suspect that, for lack of pre-testing, a high proportion of the instructional materials circulating in developing countries are incomprehensible to the people who are supposed to be learning from them.

Monitoring

Some educational efforts might start and finish in a short time, such as producing and distributing a leaflet or mounting a display stall at a one-day agricultural show. Others are ongoing activities, such as supporting correspondence students or broadcasting a weekly radio programme. For this second kind, it is a good idea to build in some method of having a look, at regular intervals, to see how it is going. You might check on the progress of all your correspondence students every three months, for example; or you might conduct a small survey of the radio audience every six months.

The purpose of this regular check is to warn you of problems which otherwise you might not have known about. For example, in a correspondence college which had many students taking many different courses and sending in their worksheets at various times, the student adviser might not notice if, say, a high proportion of the students taking mathematics were coming to a halt at lesson three. But if he received a regular report on the progress of all the students, with the information from the students' records collated and well presented, a problem like this would show up clearly. With the maths tutors and course writers he could then investigate lesson three to find out exactly why the students were stopping there and to decide what to do about it.

In time, this routine collecting of information produces a picture of the regular, normal functioning of the educational programme. This is very useful as a baseline against which to measure the effect of some innovation. For example, progress reports on correspondence students collected every three months for two years or more will give an idea of the regular dropout pattern; perhaps 40% of those who enrol are dropping out before sending in the first worksheet, a further 20% after the first worksheet but before the second, and so on. You can then use this when you launch a new course, to see if the dropout pattern on the new course is different. Or you could see if the introduction of supportive radio programmes seemed to affect the pattern.

Evaluation

I have, so far avoided the word 'evaluation' for two reasons. One is that some people seem to think that evaluation is the only use of research in distance teaching; I have tried to show in this chapter that research has several other uses. The other is that the term 'evaluation' is used to cover a range of research activities which I prefer to call by different names; some people would use the expression 'formative evaluation' to cover what I have called 'pre-testing' and 'monitoring'.

Evaluating the activities of a distance-teaching organisation means making a judgement of their value. This might include measuring the quality and efficiency of the work, seeing to what extent it is achieving its purpose, weighing up any good and bad effects of the work and deciding whether it justifies the money being spent on it. If the programme is still running, you can modify it, on the basis of the evaluation results, to try and make it work better. If the programme, such as a short-term campaign, has finished by the time the evaluation results are known, you can decide whether or not to run another one, and whether to do it differently. More generally, people in other organisations, even in other countries, can decide whether to copy your programme, and the providers of funds can decide whether to support other such programmes. It is important to realise that evaluation, even if a project has finished, is not a kind of prize-giving ceremony; the evaluation results are intended to guide somebody to making better decisions and running more successful programmes.

In my use of the term 'evaluation', it is the educational programme that is being evaluated, not the students. The evaluation of a programme might involve giving tests to the students, to see if they have learnt anything, but the purpose of this is to see how successful the programme has been. It is different from the end-of-term exams that a schoolteacher gives to his pupils where the purpose usually is to measure the achievement of each pupil.

Two examples from LDTC will illustrate the usefulness of evaluation.

We were offering a correspondence course in book-keeping, supported by a weekly radio programme. The programmes were not closely linked to the course. This was partly because our students were not all at the same place in the course, and partly because we wanted to attract a wider audience to the programmes. The programmes covered the same ground as the course, but in a different way. We evaluated the programmes by interviewing representative samples of the general public and of our own students, to find out how many people were listening to the programmes and whether they were learning anything from them. The results showed that the wider audience, even school children studying for the same book-keeping exam, were not listening with enough attention to benefit from the programmes. Our own students did find them useful, but complained about the lack of a close connection with the course. We decided that, if we were to continue with the programmes, we should have to tie them more closely to the course.

The other example concerns the booklets project. We had produced a cookery book and were intending to produce booklets on more topics. Evaluation of the cookery book showed that a very large number of housewives had bought the book and that most of them had read it, but that they had not used it as much as we had hoped they would. We decided that we should persevere with the project, since there was clearly a demand for practical booklets, but that we should also explore ways of getting the readers to make more use of the booklets, for example by encouraging women's groups to work through the booklets together at their meetings.

Appropriate research

I hope it is clear from the examples I have given that, in advocating research in distance teaching, I am not proposing massive research projects. I am not suggesting that all action should be delayed for two or three years while a research team investigates all possible topics of interest and produces a long report. While large research projects might sometimes be justified, they generally suffer from being too remote from action. By the time the report is published, the initial interest in the project has waned. Perhaps the people who originally commissioned the research have moved on and the report gets presented to someone who doesn't particularly want it. The issues will have changed in the meantime, so that the report is no longer relevant to the questions that are troubling people. And decision-makers often ignore research results anyway.

The sort of research that I am advocating is closely linked to the organisation's work. The early research throws up ideas for educational schemes. Discussion of these ideas shows the need for more research. This research then helps to further refine the ideas and so on. The researcher's task is not to produce a detailed map of the whole terrain; it is to give his colleagues a sufficiently good idea of the lie of the land to enable them to take the best course.

Many distance-teaching organisations do not have a professional, experienced researcher on the staff. Is research something that anyone can attempt or should it be left to the experts? My advice to the non-expert is to read this book (or at least the parts that are relevant to the sort of research you want to do) and have a try. There are risks, of course. You may get into difficulties and never finish the research, or you may produce results that are completely misleading. So I must quickly add a second piece of advice to the non-expert, which is not to spend too much time and money on it. But if the choice is between trying to do some research, albeit with faults, and not doing any research at all, I'd encourage you to try.

First, the possibility of doing some research puts people in a questioning frame of mind about their work and this, in itself, is valuable. The possibility of a survey of radio listeners, for example, might encourage a radio producer to think, 'I wonder what sort of people listen to my programmes? I wonder what they think of them? What exactly am I trying to get across to them anyway?' Secondly, the effort to cast a problem in a form susceptible to research often forces people to think more clearly, or in a new way, about the problem and this also has value, even if the research never gets done.

But the main value of research lies in the guidance that it gives to action, and even an incomplete and faulty piece of research can give guidance. If an artist pre-tests a drawing on some people - even just two or three people - there is a chance that he will discover flaws in the drawing and put them right. It is unlikely that, as a result of the pre-test, he will actually make the drawing worse. The results of a poorly done sample survey might be wide of the mark, but again, it is unlikely that they would be both totally misleading and sufficiently convincing to put people on a completely wrong course. Most often, the worst that can happen is that the distance-teaching staff will carry on doing what they would have done anyway. Providing, as I said, that the research is not too expensive, there is no great loss.

I have mentioned the expense of research because, of course, research costs money. If you have full-time research staff, you have to pay their salaries; you have to provide support services (office space, typing and printing, transport); occasionally you have to hire temporary assistants, for tasks such as survey interviewing; you might buy, or hire, expensive data-processing equipment. How much should you spend?

Practical research is valuable to the extent that the guidance it gives is valuable. Conducting a large sample survey of rural people, for example, is expensive, but it would be worth it if it provided basic information for designing a series of rural education projects. By contrast, it would be inappropriate to conduct an elaborate pre-test and evaluation of a single leaflet which was intended to reach only a hundred people; the research costs in this case would exceed the costs of the educational effort itself. When you finance research, you are, in effect, buying certain items of information; the price of the information is the cost of the research. You have to decide how highly you value the information before you decide how much to spend on the research. In the early days of a new organisation, or a new project, it might be justifiable to spend as much as 20% or 30% of the resources on research. When an organisation has settled down, I would think about 5% to 10% would be reasonable.

2 Before embarking on research

The chapter describes how to tackle the difficult early stages of designing a piece of research that will eventually provide useful information.

Research commissioned for the wrong reasons *Although practical research is meant to be undertaken with the purpose of providing information that will influence the work of the distance-teaching organisation (the 'action'), pieces of research are often commissioned which do not have any hope of influencing action. I describe some of the ways this can happen so that researchers can try to avoid getting into useless research projects.*

Research methods *You choose research methods that fit the problem but you also define the problem to fit your research methods. I draw a quick sketch of the methods that are described in this book.*

Defining the questions for research *It is important to get as clear an idea as possible about the information that people want from the research and about what they will do with it when they get it. I describe some ways of clarifying the questions.*

Consulting documents and experts *You may find that the information you want already exists. Experts can help in designing pieces of research.*

Time and money *Ways of estimating the cost of a piece of research and the time it will take.*

Practical research has to be linked to action. That is the purpose of doing it. But forming this link is not easy. The action problem that the research is meant to illuminate has to be clarified early on, so that a piece of research can be devised which will produce results that are relevant to that problem. Later, when the research has been completed, the results have to be fed back in such a way that the action people can and do take notice of them. These two stages - before and after the actual conduct of the research - usually get less attention than they deserve. This is unfortunate, since these are the points at which there is the greatest danger of the research and action drifting apart.

Research commissioned for the wrong reasons

Some pieces of research, though supposedly undertaken with the purpose of guiding action, do not really have any chance of influencing action. They are doomed, from the outset, to be ineffectual. There are many ways this can happen and I will describe a few so that the researcher, thus forewarned, can try to avoid such wasted effort.

One reason is, simply, vagueness. The educators feel that information of some kind would help them and they look to research to provide it, without ever thinking out clearly what information they want. If a researcher is given such a job, the best he can do is to assemble bits of information that he thinks might be useful. This information, by a lucky chance, might be exactly what the educators needed, but this is unlikely.

An example of this (not from Lesotho) concerns some research that was undertaken before a radio campaign on family planning. The educators had already decided to use radio, and they asked for background information. From the report, it looks to me as though they never took their thinking any further, so a kind of research programme was carried out without anyone deciding what it was to find out and why. Some interviewers talked to a few people in a dozen market-places around the country. They found that people in one place seemed to like film music; somewhere else they liked folk music; somewhere else they listened to the news; and somewhere else they preferred programmes in their local dialect. I cannot imagine that this was of much use to the campaign organisers. Almost certainly, it was not of enough use to justify several months of research.

People sometimes commission research without having any action problem in mind. Perhaps they simply like the idea of having some research done; they feel that it confers prestige. Or perhaps they think it would be interesting just to find out about something. If the research is not directed towards an action problem, not even vaguely, then the results are unlikely to influence action.

Another misuse of research is when people have a decision to make but, for some reason, they are reluctant to make it. They often call for more research into the topic, not because they really need more information to help make the decision, but just to postpone the decision making.

Probably the most common cause of ineffectual research, however, is when the action, which the research is supposed to be influencing, is not in fact open to influence. A project director might request an evaluation of the project even though he has no intention of modifying the project in the light of the results. Perhaps there is not even any possibility of making significant changes to the project. He is not looking to research to provide guidance; he is seeking reassurance. Or someone might commission research to provide support for his point of view in a policy debate. Again, he is not intending to be guided by the research; he just wants it to confirm something he already believes.

One might argue that research, even though commissioned for the wrong reasons, will sometimes throw up findings that force people to change their minds - the people did not want guidance, perhaps, but the research has clear implications which they cannot ignore. It would be nice if this was true, but in my experience it isn't. People who have made up their minds are inflexible. Put yourself in the shoes of a policy-maker who is faced with a research result which suggests that he should change his policies. Either you can accept the research results and change your policies or you can keep your policies and ignore the research results. The second course is much easier.

The Lesotho Distance Teaching Centre's commitment to radio provides an example of this. International Extension College, who initiated LDTC, had coined the expression 'three-way teaching' to describe the combination of correspondence courses, broadcasting and face-to-face teaching that they advocated, and LDTC was established to explore the use of any available media for distance-teaching, so it was part of the original plan that LDTC would use radio. Early research suggested that radio ownership was fairly low (about 17% of households had radios in 1975) and that reception of the national radio station was poor. Nonetheless, when the first students had been enrolled for correspondence courses, several sets of radio programmes were written, produced and broadcast to accompany the courses. More than one evaluation survey over the next two years reported that the audience for these programmes was very small, even among LDTC's own correspondence students. If research had influenced action in any straightforward way, LDTC would have reduced its radio work. But the commitment to radio remained and indeed increased. The radio section was expanded; more programmes were produced; an adviser on radio was recruited from overseas, and eventually a whole studio was built and equipped. People who like an idea are not easily put off it.*

Research methods

You might think that the logical way to design a piece of practical research would be to define the problem and then to select the most appropriate research method to tackle the problem. What happens in practice is a bit more complicated. You begin with an idea of the range of research methods at your disposal and, as soon as you are given the problem, you start thinking how you might tackle it. As a research design begins to form in your mind, you see the ways in which the problem needs to be defined. You develop a research design to fit the problem, but you also define the problem to fit the research design. The two parts interact.

In this section, I draw a quick sketch of the research methods that I am going to describe in more detail later in the book. Then I suggest ways in which, with these methods in mind, you can define a problem so as to make it susceptible to research.

---

* In all fairness I should mention some of the arguments that were put forward for building up the radio section. Radio Lesotho had plans to increase its transmitting power; radio ownership was likely to go up rather than down; and it was part of LDTC's job to exploit the educational potential of radio, so it should continue this work even though the early results had been disappointing. But I think the example still illustrates my point: research results, even when they are clear, repeated and not in dispute, do not necessarily influence policy in the way you might expect.

Observation is a research method. Before designing materials to teach basic numeracy at LDTC, we wanted to know what sort of calculations people were called upon to make in their everyday lives. We spent some time in village shops noting down the purchases that people made. We found that three-fifths of the shoppers bought just one item and that four-fifths of them spent less than 50 cents. To get an idea of the uses of literacy in Lesotho, we catalogued all the reading matter in a number of rural homes (with the owners' permission, of course). You could learn a lot about rural life from observing fields, gardens, crops, livestock, farm implements and so on.

Consulting records is another research method. LDTC used the official statistics on road accidents and traffic offences to assess the impact of a road safety campaign. In that case, the records were collected by another agency as part of its regular operations. You can, of course, collect records of your own. A correspondence college, for instance, generates a set of records in keeping track of its students' progress.

Talking to people is, obviously, a basic research method. Social scientists distinguish broadly between two ways of doing it. The first way is that you hold something like an ordinary conversation, either with an individual person or with a group. You might ask certain questions to guide the conversation on to the topics that you are interested in, but the people are free to tell you anything they want to tell you in their own way. The other way is that you interview people with a questionnaire, reading out the questions exactly as they are written and recording the answers in a systematic way. For example, if you wanted to know how many farmers owned oxen, you might take a sample of farmers and put the same question to each one, 'Do you own any oxen?' Then you add up the answers to find, say, that 20% own oxen. This second method is known as a social survey. A variation on this is to give people the questionnaire, or to send it through the post, and get them to fill in the answers themselves.

Experiments can also be used in social research. For example, if you wanted to know whether long letters of encouragement had any effect on correspondence students, you might arrange for some students to receive long letters and other students to receive short ones, and you would see if it made any difference. A special kind of experiment is when you investigate the feasibility of some idea by actually putting it into practice, perhaps on a limited scale, and seeing how it goes. This is sometimes called 'action research'. For instance, to find out if there was any demand for a radio magazine programme for housewives, you might broadcast one for a few weeks, to see what response it received.

There are other methods apart from these. Later in the book I will describe some specialised techniques for assessing instructional materials, for example. But these are the basic ones.

Different methods provide different sorts of information. In an experiment, you alter things in some way in order to see what happens. With the other methods, you don't alter things, or at least you try not to; rather you try to get a picture of the way things are. Methods which involve counting and calculating provide a picture in figures; they can answer questions of the form 'How many ..... ?' 'What proportion of ..... ?' 'What is the average ..... ?' Methods which don't involve counting, especially conversation and group discussion, provide a picture rather in words or images.

Defining the questions for research

When people want some research on a problem, they often present the problem in a way that is not immediately susceptible to research. They tend to use vague phrases like 'investigate the feasibility of' or 'assess the effectiveness of'. But if there is to be any hope that they will use the results when they get them, they have to say more precisely what sort of results they want.

The first step in defining the research question is to ask them 'What do you want to find out?' Unfortunately someone who has begun with a vague phrase is likely to continue with more vague phrases. A group of educators might request information on 'community involvement in education', for example. If you ask them what they want to know, they are likely to make more statements, using different words but equally vague - 'the concept of community-based schooling', 'the whole area of village perceptions of the formal school system', 'what goes on across the interface between school and community' and so on. People like talking in this way. It sounds good and it saves them the effort of thinking. But if it is not clarified, it leads to poor research. If the researcher doesn't know what the questions are, how can he find the answers?

Another ploy is to ask people to guess what the results of the research might be. This forces them to think of what they might expect to get out of the research, and their answers might reveal more precisely the questions that they are interested in. I gave the example earlier of a vague piece of research that was conducted in response to a vague request for 'background information' about radio. If the campaign organisers had been asked,'What results do you expect?' they might possibly have given answers like these:

> 'I expect that only a very few people own radios.'
> 'I would guess that the news is the peak listening time.'
> 'I think people only listen for entertainment; if it sounds like serious talk, they will switch off.'
> 'I'm worried if only the rich have radios; it's the poor we really want to reach.'

Statements like these are a great improvement on a vague request for background information. One could begin to frame a research project around them, to find out what proportion of people own radios, how radio-owners differ from other people, which programmes people listen to most, at what times most people listen, and so on.

If that fails, you have to make your own suggestions about questions that the research might tackle, making use of any hints you can pick up from their vague statements. To the people who were interested in 'community involvement in education', you might offer ideas along these lines: 'We can find out about the membership of school boards from records at the Ministry of Education. We can talk to a few teachers to see if they welcome parents' interest in the children's education. We could do an interview survey of parents to find out how often they visit their children's schools. Is this the kind of thing you want?'

If, despite your efforts, the research questions remain vague, you can sometimes do a quick piece of research to clarify the questions. This is known as 'exploratory research'. To continue the same example, you might visit a village school and talk to some teachers, in a very general way, about their contacts with parents and any other aspects of the relationship between the school and the village. You might then be able to identify more precisely the topics for more detailed research.

When you have made the research questions sufficiently specific, you can then put another important question to the people who are commissioning the research - 'What will you do if we find that ... ?' (you insert your own guess of what the result will be). It is important to get people thinking about this early on. People who commission a piece of research are often unable to visualise what the results will look like. The danger is that, when they eventually get the results, they won't know what to do with them.

If people force themselves to face this question, they sometimes realise that the research results will make no difference. They have asked for research on a particular course of action, but they are not seriously considering any alternative course of action; perhaps they cannot even imagine an alternative course of action. So the research is not really needed at all.

Thinking about this question also prepares people for the possibility that the research results might not be decisive. Suppose that two people are preparing literacy materials for out-of-school children; they have different opinions about whether to expect parents to help their children with the materials. They will probably express their opinions in sentences like, 'I'm sure parents will take this opportunity to help their children,' or 'I don't think we can rely on their help.' They commission a piece of research to find out whether or not parents help their children with reading and writing. Almost certainly, the research will find that some do and some don't. The result of the research will be a percentage, such as '40% (or 20% or 60%) of parents have given their children some help with reading and writing, and 15% (or 5% or 30%) have done so in the last month.' If the literacy people can think, in advance, about what they would do with such a result, there is some hope that they will actually use the result when it comes. But if they commission the research in the hope of getting a 'Yes/No' answer to their question and then they eventually get a percentage instead, they may not know what to do with it.

Since part of the purpose of this book is to convince people of the usefulness of research in distance teaching, the point I am about to make might seem out of character. But there is a danger of thinking up so many questions that seem to require research that you prevent yourself from taking any action at all. Consider an organisation that is wondering about publishing an instructional booklet. What would be the best topic? How long should the booklet be? What style should it be written in? Should it use illustrations? What sort of illustrations? Should it use colour? Should it be sold or given out free? If sold, at what price? And so on. You could spend years doing preparatory research for this one booklet.

You cannot research everything. In fact, of all the possible questions that you might think up, you can only research a few. Two pieces of advice follow from this. First, select those questions on which research can provide the most help. Second, don't forget action research. If you wanted to find out if people would buy a booklet priced at 20 cents, the best research method would be to offer booklets at 20 cents and see if people bought them. Research does not necessarily have to precede action. An alert researcher, by arranging to gather data from an action project, can use the action itself as a research method.

Consulting documents and experts

In academic research, when you are hoping to contribute to knowledge on some topic of general interest, it is well established practice to review previous work. For example, if you were investigating whether children of a certain age can or cannot grasp certain scientific concepts, it would be foolish to proceed with a research project without having consulted the substantial literature that already exists on this topic.

In practical research, the questions are usually more specific to the place, time and project that you are working on. What are the main cattle diseases in this country? What beliefs do these people have about types of food? How good is radio reception in the northern districts? So there is not so much relevant literature that you have to consult. Nonetheless, even in a small country like Lesotho, government departments and other agencies carry out many pieces of research, so you may find that someone else has already obtained the facts you want. At LDTC, for example, before we wrote a booklet on vegetable growing, we were able to use figures from the Ministry of Agriculture on the proportion of people who grow vegetables and on the types of vegetables they grow. When preparing a booklet on child care, we were able to use reports from two health education projects on traditional child care practices.

If the problem is about the design of distance-teaching materials, or the organisation of a distance-teaching programme, the experience of other countries might be relevant. There are various agencies you can write to for this kind of information; some are listed in Appendix 6. Try to describe precisely the problem you have or the sort of information you feel you need, perhaps enclosing background papers on your project. This will help them to locate material that is really relevant to your research. There is a possibility that the people who answer your letter may not know any more than you do about work that has been done that is relevant to your problem, so they may just refer you to another agency or send you publications that they happen to have, which might be quite irrelevant. However, they do sometimes produce something useful, so it is worth making the effort to consult them.

In addition to international agencies, it is useful to establish contacts with colleagues in similar organisations in other countries. People tend to be more honest in informal letters or telephone calls than in formal reports, particularly about mistakes they have made or awkward problems they have encountered.

On technical matters connected with research design, you might be able to consult experts in research and statistics, perhaps in the government statistics department or in the university. If you do need expert advice, it is better to seek it early on. If you embark on a badly designed piece of research and gather a pile of faulty data, there is not much that an expert can do at a late stage to put it right.

A word of warning about experts, especially statisticians. When consulted for a professional opinion, they tend, naturally, to be conscious of their professional status. They do not want to condone a research design which another expert might consider shoddy. So they are inclined to give you advice that comes straight out of the textbooks. This can lead them to recommend a research design that is more ambitious than you can handle and more refined than it needs to be for your purposes. I am not suggesting that you ignore what they say. But you should be prepared to question their recommendations and discuss the research design with them, rather than just accept what they tell you.

Time and money

Educators generally need information fairly quickly. If a new college requests a survey of correspondence students to help it decide which subjects to offer first, it does not want to wait two years for the results; it needs them in a month or two. If a course writer wants to know whether to adopt this style or that style, he needs to know before he has written too much of the course.

It is important to be realistic about this. A social survey of the kind described in later chapters will take at least two months and probably much longer. If this means that the results will arrive too late, it may be better not to embark on it at all. You might decide not to do any research on the topic, or you might select a research method that will produce results more quickly. A quick piece of research will probably produce results which are less reliable, less academically respectable, than a larger one. But rough results at the right time are much more use than elegant results that arrive too late.

As well as being short of time, the research department is often short of money. Here again, it is important to be realistic. A piece of research has to reach the stage of producing some results before it can hope to have much effect on action. If it is abandoned, for lack of money, at any point before that stage is reached, the money already spent on it will have been largely wasted. So the researcher must be confident that the research he has designed can be completed with the funds available for it.

In order to design a piece of research with full regard to time and money, it is of course necessary to be able to estimate how much time a piece of research will take and how much money it will cost. This is not easy. To estimate the time, it is best to divide the research into its different stages and to make an estimate for each stage. For an interview survey, for example, you might divide the research into questionnaire design, recruiting and training interviewers, fieldwork, data processing, analysis of results, writing report. You estimate the time that each stage will take and then you add them up. Similarly, in estimating the money it will cost, try to think of all the separate items that will cost money, e.g. salaries of research staff, travel expenses, use of data-processing equipment, typing and printing of final report.

If you have little experience of research, ask the opinions of people who have done it before and revise your estimates accordingly. Then, when you have made the best estimate you can make, multiply it by 1.5 or even by 2. This might seem ridiculous, but people's estimates (mine included) are almost always too optimistic. This is partly because you do not want a piece of research to take too much time and cost too much money, so you persuade yourself that it won't. And it is partly because you assume that things will go smoothly, whereas in fact they never do.

With more experience you will get better at making these estimates. Keep a record of the time and money that your pieces of research actually do take. This will provide the most reliable basis for making estimates of future research projects.

3 Observation and discussion

Observation is a good research method if you want to give yourself a picture of people, places or events. To get the most out of an observation visit, make notes in advance and write a report as soon as possible afterwards. Participant observation.

Reactivity People who are aware that they are being observed behave in a different way from usual. There are ways to reduce this problem.

Discussion Discussing things with people, perhaps in groups, can reveal attitudes which the researcher had not expected. The interviewer's approach to people is more important than his knowledge of the topic under discussion.

Recording discussions You can take notes, during or after a discussion, or you can use a tape-recorder. Beware of generating more material than you can analyse.

Analysis Rearrange the material to bring out the main themes. Bear in mind the possible influence that the interviewer had on the discussion. Avoid the temptation to treat the results as statistical data.

Observation

People who would not regard themselves as researchers use observation as the natural way of finding out about things. Say you were designing materials for health education lectures in rural clinics, and you wanted to get an idea of the situation in which the materials would be used; the obvious thing to do would be to go and take a look at some rural clinics.

If you cannot picture clearly the sort of people who will be learning from a distance-teaching project, or the setting in which they will be doing it, observation is a good method to use early on. When I arrived in Lesotho, to initiate distance-teaching projects for rural people, I had very little idea of what rural life in Lesotho was like. So I spent about six weeks living in three different villages. I kept a daily diary and the following extract gives an idea of the sort of things I learnt from this experience:

> The household's herdboy was late setting out this morning. Sekonyela [my assistant] says the herdboy had taken the cattle out in the night - apparently you sometimes do this to steal a good bit of pasture that's reserved for someone else. We went with the herdboy. He lent me his stick. Solid thing. Decorated. The pasture was only five minutes away. Four or five other herds were already there.

A herdboy wears a loincloth, a blanket (old and tatty), wellington boots and some kind of hat. You move your herd slowly round the pasture with whistles, shouts and clods of earth thrown at them. All the cows have names. Sekonyela says there's no danger of confusion. The boys gather in little groups around midday, sit down, chat, play moraba-raba (a sort of board game that you play on the ground or on a rock) or fight in a playful way with their sticks. One boy had a lesiba, a musical instrument. It's a straight stick about a yard long, with a feather quill at one end. You put your lips to the quill and blow, or suck, or hum (I'm not sure which - perhaps all three). Clearly not easy and it makes a rather dull noise - just three or four notes.

Tried a bit of bird-catching. The idea is that you surprise them from their nests in the grass and then hit them in flight with your stick. Seemed unlikely to me and we got nowhere near catching any.

When I went to stay in these villages I did not have any particular questions that I wanted to find the answers to. The purpose was, rather, for me to build up a picture of rural people and the lives they led, since they were the people who were intended to benefit, eventually, from our distance-teaching efforts. It is important, in distance teaching, that the picture you have in your mind of the people who will be receiving your materials should be a reasonably accurate one. To take the example of rural education, it is all too easy, sitting in an office in a town, to write materials for rural people which give advice such as 'To destroy this pest, use Thiodan,' or 'Do not move a person with a broken leg; fetch a doctor,' or 'Eat fresh fruit every day.' If the village store does not stock Thiodan, or if the nearest doctor is fifty miles away, or if fresh fruit is only available for two months of the year, these pieces of advice are pretty useless. A writer would be less likely to make such mistakes if he went to live in a village for a while, even if only for a week or two. Although this obviously applies to people recruited from other countries, it can also apply to local staff; someone who has lived in the city all his life, or even a countryman who left his village when he went to secondary school many years ago, can be just as out of touch with rural life.

A more specific example of the use of observation is a visit we made to a village meeting, held outdoors, where a fieldworker was speaking on family planning. We had been commissioned to produce materials to help the fieldworkers give these lectures. One simple discovery we made was that some members of the audience sat as much as forty metres away from the speaker; so any visual aids we might design would have to be big and clear. Another discovery was that some members of the audience got very angry. One man, waving a heavy walking-stick, even threatened to attack the speaker. At the time, I couldn't understand what people were getting so angry about. Obviously we needed to find out more about people's attitudes to family planning and we discovered, eventually, that their anger was connected with their beliefs about breast-feeding (see page 12).

After an observation visit, you should make notes as soon as possible. This forces you to think about what you have just seen. It also enables you to record a lot of detail which, otherwise, you tend to forget.

If you are going out specifically to observe something, it is useful to prepare a short checklist of points to look for. For example, if you were designing a training course for the committee members of co-operatives, you might attend a meeting of a co-op committee and prepare yourself in advance with notes like these:

Number present? Was there a chairman?
Did they have an agenda? Did anyone
record the decisions?

If you know even more precisely what you want to observe, you can prepare a recording sheet to help you take down the details you want. In the last chapter I mentioned that we observed the purchases that people made in village shops. Since we knew, in advance, exactly what details we wanted, we could prepare a recording sheet. The researcher, sitting in a corner of the shop, just noted the details, one line for each customer:

Approximate age of customer	Number of items bought	Total cost	Money offered	Change received

Sometimes it is useful for the researcher to participate in some activity alongside the people he is studying; this is called 'participant observation'. A colleague in Botswana, as part of the evaluation of a radio learning group campaign, arranged for observers to attend the meetings of various radio learning groups, to listen with them to the radio programmes and to join in the discussion afterwards.

Taking this idea a step further, the best way to find out about the problems of doing something is to try doing it. To learn something, at first hand, about the problems of farming in Lesotho, I sharecropped a small field of beans one year, with a great deal of help from my assistant. I learnt a lot from this about the problems that ordinary farmers had, and about the hazards of sharecropping in particular. In a similar spirit, a colleague enrolled for a correspondence course, mainly because he wanted to study that course, but also to discover for himself the difficulties of studying by correspondence. Evidently he did discover the difficulties; like many correspondence students, he failed to complete the course.

Reactivity

Any method of studying people, including observation, encounters what social scientists call the problem of reactivity; people react to being studied. You are trying to study people's natural, ordinary behaviour, but they know they are being studied so they behave differently from usual. They put on a special show, so to speak. LDTC staff wanted to observe an ordinary meeting of a women's group using the How to crochet booklet, but they were unable to do so; when they visited a village, their presence attracted an exceptionally large number of women to the meeting. It was impossible for them to observe a completely ordinary meeting since any meeting they attended was no longer an ordinary one.

This problem is at its worst when the observer is obviously an outsider, such as a white expatriate in a Lesotho village. So you can reduce the problem by selecting an observer who will fit easily into the situation. Also the observer should avoid drawing attention to himself. One observer, arriving at a village on foot and without equipment, would be less conspicuous than a team arriving in a Landrover with cameras, tape-recorders and notepads. Another way to counter this problem is to stay around long enough for people to get used to you; someone I met in Lesotho who was studying primary education used to stay at a school for several days, so that the staff and children would get accustomed to his presence and resume their normal routine.

The problem would not arise at all, of course, if the people did not know that they were being studied, so the solution might be to observe people without their knowledge; perhaps the observer could hide or disguise himself. But this clearly raises an ethical question; is it right to study people without their knowledge? Opinions vary widely on this. My opinion is that in general you shouldn't do it, but that there are some circumstances in which it is justified. Rather than discuss it in general terms here, however, I will return to this question later, in connection with a particular example in evaluation (page 217).

Discussion

Like observation, discussion is a natural method of finding things out - if you want to know what cattle-farmers think about the problems of cattle-farming, the obvious thing to do is to go and talk to some cattle-farmers. A discussion conducted for research purposes is very like an ordinary conversation. The researcher asks questions in order to guide the discussion on to the topics that he is interested in, but the person he is talking to is free to express his opinions in his own way, to make points that the researcher had perhaps not expected, or even to raise new topics that the researcher had not thought of and would not have asked about. A researcher can also hold a discussion with a group of people. This can be more valuable sometimes, because the discussion between the group members themselves throws up further points.

The great advantage of discussion is its openness. The researcher, of course, has his own view of the topics, but other people might have a very different view. By allowing people to say what they want, the researcher can begin to see the topics through their eyes. For example, a rural development project asked LDTC to design training materials for their village agents. These were villagers who kept stocks of the project's seed and fertiliser and sold them to farmers, receiving a small commission on each sale. Many of the agents were not coping well with the record-keeping system that the project had designed for them; this was why the project had asked LDTC to help. Before designing the materials, we gathered several of these village agents together and held a discussion with them. What emerged clearly was that the agents did not think that the record-keeping was a major problem; they were far more worried about the security of the stocks and the cash that they held for the project, a point to which the project had not paid much attention. We were able to tell the project that the agents were not likely to work hard at improving their record-keeping, which was more for the project's benefit than for their own, unless the project paid more attention to the agents' worries.

The success of a discussion depends greatly on the person conducting it. (I'll call him 'the interviewer'.) The person in charge of the research is not necessarily the right person to conduct the discussion. As part of our preparations for a book about baby care, we wanted to discuss childbirth practices with a group of village women. Obviously, I myself was not the right person to conduct the discussion, since I did not speak the language well enough. Besides, the village women would not have discussed these matters with a man. Our usual fieldworkers were also unsuitable, since they were young, unmarried women and the village women would not have thought it proper to discuss the details of childbirth with them either. The interviewer had to be a married woman with children of her own. Only then would the village women feel at ease.

The interviewer needs to know something about the topic under research. Then he (or she) can ask questions that are relevant to it. But the interviewer does not need to be an expert on the topic. It is more important that he should have the right approach to people. He must be able to put people at ease, to stimulate the discussion, to guide it gently so that it does not stray too far from the topics in which he is interested, to recognise if a new point is important and to follow it up, to encourage the more diffident members to express their views, and generally to direct the discussion without dominating it.

The interviewer's own contributions to the discussion should be open and non-committal. He should try not to express his own opinions, since this affects what people say. Most people are polite and deferential and they would not want to disagree with the interviewer. On the other hand, some people are argumentative and would take an opposing opinion just to provoke a debate with the interviewer. Either way, people would not be expressing their true opinions. The interviewer's job is not to take part in the discussion like a group member, but to prompt and clarify the group's discussion. So he should introduce the topics in a general way to start the discussion and thereafter keep it going with questions such as, 'Can you tell me more about that?' 'I wonder how the other members react to that?' 'How do you mean exactly?' From time to time he might try to summarise what people have said, to clarify it and to make sure that he has understood them correctly.

As with the observation visit, it is useful to prepare for a discussion by jotting down the questions you want to raise. This is not a questionnaire from which you will read the questions one by one. It is rather a reminder of points you want to cover in the discussion. A discussion of this type, when conducted with just one or two people, is sometimes called a 'semi-structured interview', meaning that it is half-way between a completely informal conversation and a questionnaire interview.

Recording discussions

Conducting a group discussion requires constant attention to what is being said, so it is difficult for an interviewer to make notes during the discussion. One way round this problem is to have two interviewers. One can take notes (if this does not upset the group members) while the other conducts the discussion. One can raise points that the other has missed, and one can remind the other afterwards of things that were said, when they write up their report of the discussion.

But it is not always possible to have two interviewers. Besides, the group members may be inhibited if they can see that their remarks are being written down. If there is no alternative, the interviewer just has to try to remember what is said. If so, he has to write notes as soon as possible after the meeting. But this is not satisfactory since he will, inevitably, forget some things. Another possibility is to use a tape-recorder.

To tape-record a group discussion, you need a good quality tape-recorder and microphone. You should practise with it in advance of the discussion, so that you are familiar with the controls and so that you know where best to place the microphone. (Some modern cassette recorders have good, built-in microphones. These are easier to use and may be less threatening to the group members.) Make sure you have enough tapes and batteries before you start. Don't attempt to record people without their knowledge. Draw attention to the tape-recorder early on, explain why you want to use it and make sure everyone is happy about it before you switch it on.

In the next few days after the discussion, listen to the recording and make notes on the main points. If the voices are clear and you feel you want a complete transcript, you can ask a typist to type out the discussion in full. If the recording is not clear enough, or if the typist just cannot do it (it is quite difficult typing from real speech), you have to edit it. That is, you listen to the recording and decide which bits you want to have typed, and then you write these out (or dictate them clearly into another tape-recorder) for typing. You may want to edit the tape anyway, especially if the discussion was long and repetitive.

If you use a tape-recorder, use it sparingly. Although it saves you the effort of taking notes during the discussion, it creates more work in the long run. Simply listening to a tape of a discussion takes as long, obviously, as the original discussion. Listening to it and stopping it to make notes takes longer. Typing it out in full takes a lot longer, and a complete typewritten transcript of an hour's discussion fills many sheets of paper. After just three or four discussions, you are likely to find that you have more tape and paper than you can handle.

Analysis

Since the purpose of an open discussion is to allow people to say whatever they want, no two discussions will be the same, even if they are on the same general topics. So it is not easy to organise and condense the results. First, read through all the typescripts and list the main themes. Then go through the typescripts again and mark on them where the same theme crops up in different discussions, or in different parts of the same discussion. Then cut up the typescripts with scissors, and paste together on a large sheet of paper those sections on the same theme. Then, with the different sections on a single theme in front of you, try to summarise in your own words the various opinions that were expressed and repeat this for each of the main themes.

When interpreting the results of discussions, try to assess how much influence the interviewer had. Because the interviewer plays such an important part in the discussion, he can have a great influence on what is said, and he may not even be aware of this himself. Suppose that a woman conducted a group discussion with some village women on the subject of family planning. Imagine that the interviewer, educated at a Catholic secondary school, has given a lot of thought, as regards her own life, to the Church's teaching on family planning. She might simply assume that the Church has an important influence on people's attitudes because it has had a large influence on her own. In the group discussion she would be inclined to raise this aspect of the topic and to get the group members talking about it. So the results would suggest that the Church has a strong influence on these women's attitudes to family planning. But this might be completely wrong. Perhaps the women do not pay great attention to the Church's teaching on this matter; to a different interviewer they might never have mentioned it.

Avoid the temptation to treat the results of discussions as statistical data. That is, do not draw conclusions such as 'Most people hold this opinion.' The sample of people who have taken part in the discussions is too small and almost certainly not representative. (I'll explain in the next chapter what I mean by this.) In addition, the results could have been strongly influenced by just one or two leading members of the group.

Discussions are good for giving you an insight into people's thoughts and beliefs. They can show you how things are linked together in people's minds; they can suggest aspects of a topic that you had never thought of. For this reason they are particularly suitable for exploratory research - research undertaken to clarify the questions for further research. It is often useful, for instance, to hold a group discussion on a topic before you try to write a questionnaire on it. The results of discussions are ideas, patterns of opinions and attitudes, described in words. If, however, you want results from which you can calculate statistics, i.e. results described in figures, you have to use different research methods.

Social surveys: introduction to chapters 4–9

The basic idea of a social survey is simple. Suppose you wanted to know how many households possessed a radio. You might visit, say, two hundred heads of households and say to each one, 'Do you have a radio?' You record their answers on slips of paper and then count them up at the end. You find that, say, 50 out of the 200 said Yes, so you conclude that about a quarter of all households have a radio.

Two basic concepts are contained in this example. The first is that you do not have to visit every household in the country; it is enough just to visit a sample of households, provided that the sample is representative. That is to say, your 200 households give you a picture in miniature of all the households in the country. To get a sample that is representative, you have to choose your households carefully. This is known as sampling.

The second concept is that you add together the answers to provide a summary, in numbers, of what you have found. It is essential, therefore, that everyone gets asked exactly the same question. If some people were asked, 'Do you have a radio?' and other people were asked, 'Do you ever listen to the radio?' it would be absurd to add their answers together. To ensure that everyone gets the same question, you write a questionnaire.

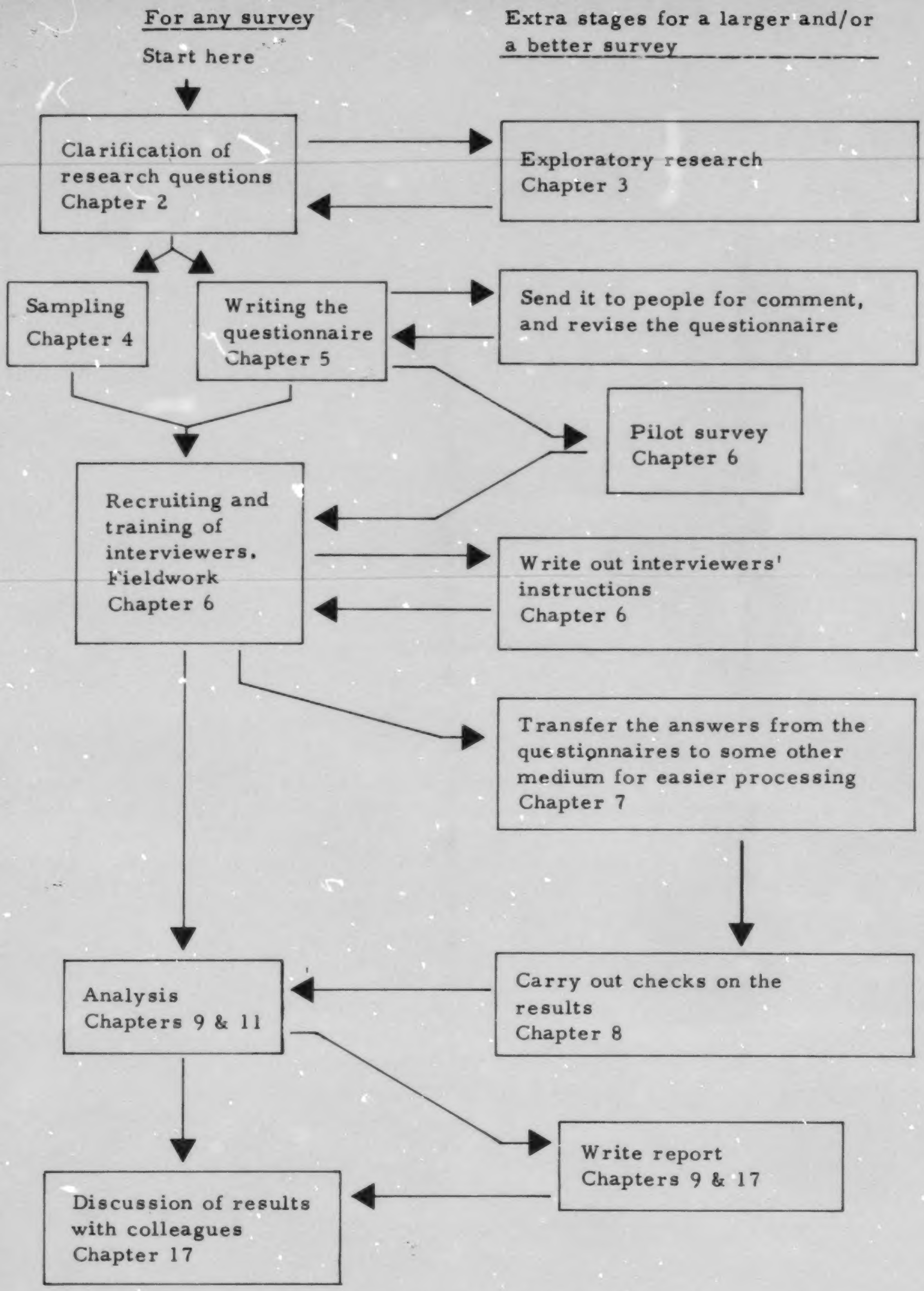
Social surveys provide quantitative results, i.e. results presented in expressions like '46%', 'three-fifths', 'nearly all', 'most', 'very few'. Some people become nervous when they see a number; 'I don't understand statistics,' they say. And even people who do understand statistics have to admit that a report containing lots of quantitative results makes heavy reading. But information about large groups of people has to be in this form. 'How many households have radios?' or 'What proportion of our students complete their courses?' are obviously questions that require quantitative answers. And questions which are not, at first sight, of this form often turn out to require quantitative answers. 'How well is our campaign message getting across?' really means 'How many people have heard our radio spots, or seen our posters, or read our leaflets, and what proportion can remember the campaign slogan?'

Surveys vary greatly in scale. Most of the surveys we did at LDTC took samples of 50-100 people, though occasionally we took a sample of 300-400. Major surveys conducted by government departments may take samples of several thousand people. At the other end of the scale, taking a draft pamphlet out to 20 farmers and asking them some questions about it is also a social survey - a very small, specialised one. Some of the techniques you use depend on the scale of the survey - a simple method of data processing that might work for a small survey would not work for a large one, for instance - but the basic principles apply to all surveys regardless of their size.

The diagram on the next page shows the main stages of a survey. The stages down the right-hand side of the diagram are optional, in the sense that it is possible to do a small, rough-and-ready survey without them. Suppose you were pre-testing a pamphlet. You would first make sure that you understood what the pamphlet-writer wanted to get across (clarification of research questions). You would decide what sort of people you were going to show it to and how you were going to select them (sampling) and you would write some questions to see if they had understood the pamphlet (writing the questionnaire). You would select some people to conduct the interviews and you would explain how they were to do it (recruiting and training of interviewers) and then they would carry out the interviews (fieldwork). You would collect up the questionnaires, inspect the answers and perhaps count them up and express the results in a few tables (analysis). Finally you would discuss the results with the pamphlet-writer so that he could decide what improvements to make to his pamphlet. The whole operation might be done inside a week.

By contrast you might decide to do a survey of several hundred people in order to gather a large amount of detailed information on an important topic. At LDTC we conducted three surveys of this size in the first three years, one of which was the survey of rural people's reading abilities that I mentioned in Chapter 1. For a survey of that kind, which would be an ambitious undertaking for a small organisation, you would include some or all of the stages down the right-hand side of the diagram, and the complete survey could take about 18 months or two years. And there are intermediate points between these extremes. Before writing a booklet at LDTC about making clothes (sewing, knitting, etc.) we did a small survey to find out how many rural women made clothes at home and what level of advice they would find useful from a booklet. We used a straightforward questionnaire, interviewed about 100 women, did some simple analysis and wrote a short report; the whole survey took about three months.

I have not attempted to write a textbook on survey methods, but Chapters 4 to 9, along with parts of Chapters 2, 3, 11 and 17, tell you as much as you need to know in order to conduct surveys in the range I've just described - from a one-week's test of a pamphlet to a two-year survey of several hundred people. If you only want to do small, quick surveys, you don't have to read all these chapters, but I recommend that you do if you have the time and the interest, on the grounds that, even if you decide to skip certain stages, you ought to know what it is that you're skipping.


For any survey
Start here
Extra stages for a larger and/or a better survey
Clarification of research questions Chapter 2
Exploratory research Chapter 3
Sampling Chapter 4
Writing the questionnaire Chapter 5
Send it to people for comment, and revise the questionnaire
Pilot survey Chapter 6
Recruiting and training of interviewers. Fieldwork Chapter 6
Write out interviewers' instructions Chapter 6
Transfer the answers from the questionnaires to some other medium for easier processing Chapter 7
Carry out checks on the results Chapter 8
Analysis Chapters 9 & 11
Write report Chapters 9 & 17
Discussion of results with colleagues Chapter 17

The main stages of a social survey

4 Sampling

Who to interview? Decide who are the most appropriate sort of people to interview and take a sample of them. Be suspicious of secondhand information.

Using randomness to achieve a representative sample You want information about a very large group of people but you interview only a sample of them. You want the sample to be representative of the large group. You use random selection to achieve this.

Simple random samples and systematic random samples How to use a table of random numbers to draw a random sample.

Stratified samples Sampling separately from each stratum of the population makes a sample more representative.

Multi-stage sampling (cluster samples) You draw a sample of 'clusters' of people (e.g. villages) and then a sample of people within each selected cluster.

Heads of households in rural Lesotho An example of how we might draw a sample for a survey conducted by LDTC.

How many interviews? It is more important to have a good sample than to have a large one.

Who to interview?

First, you have to decide what kind of people you want to interview. Obviously, this depends on the topic of the survey. If you wanted to find out what proportion of households owned a radio, you might decide to interview heads of households. If you wanted to find out how mothers were feeding their infants, you might interview mothers of children under five.

It is worth taking some trouble to interview the right people. For example, if you are finding out people's attitudes to nutrition before a national campaign on nutrition, and if the principal audience for this campaign is going to be rural people, then get out to the villages and interview rural people. Don't just interview people in and around the capital city. When pre-testing a family planning pamphlet at LDTC, we found that people who lived in a village on the edge of the capital had attitudes quite different from those who lived 50 km away.

Be suspicious about secondhand information. If you want to know something about a certain sort of people, try to interview some of those people; don't be tempted to get your information from some other people instead. An example will show what I mean. The first booklet that we published at LDTC was a cookery book produced in collaboration with the Catholic Relief Services (CRS). Its Sesotho title was Buka ea CRS ea ho Apeha. It occurred to me that, since 'CRS' was the abbreviation of the organisation's English name, the initials might be unfamiliar to the rural women of Lesotho, most of whom did not speak English. I put this question to the nurses who worked for CRS and they assured me that people throughout the country would be familiar with these initials. So we left the title as it was. We discovered, much later, in evaluating the booklet, that most of the women who bought the booklet did not recognise the initials 'CRS'. So far as they were concerned, this part of the title was meaningless. The fault was not with the CRS nurses; they gave me their honest opinion, which happened to be wrong. The fault was mine for being lazy; I should have asked some rural women.

I am not saying that you should never collect secondhand information, but that you should not accept it as a substitute for firsthand information. For example, if you wanted to know whether secondary school dropouts were interested in continuing their education by correspondence, you should ask some secondary school dropouts. You might also ask head teachers for their opinions on the needs of school dropouts - and their opinions might be interesting - but you should not interview headteachers as a substitute for interviewing dropouts themselves.

Using randomness to achieve a representative sample

If you are interested in only a small group of people, you may be able to interview all of them. Suppose that you wanted to interview dentists about the need for a campaign on dental health; if there were not many dentists in the country, you might interview them all. Generally, however, there are far too many people of the kind you are interested in. In Lesotho, there are about 200 000 heads of households; if we wanted to find out something about, say, their poultry-keeping, we obviously could not visit all of them. In such cases you have to take a sample.

A sample is intended to be representative. You do not have the time and resources to interview 200 000 people; you interview, say, only 100. These 100 people are intended to represent the 200 000. If you find that 50 of the 100 people own poultry, you conclude that about half of the 200 000 people own poultry.

It is obviously very important to select a sample that is representative. If you did all your 100 interviews in one village that had recently been attacked by a poultry disease, you might find that only ten of the households had poultry. This would give a most misleading picture of the nation as a whole. How do you select a representative sample?

Let's say you are investigating the possible use of distance teaching in educating the committee members of co-operative societies. You decide to do a survey of the co-ops. There are 700 co-operative societies in the country and you decide that you can visit 50. You could choose those co-ops that were nearest to your office, but it is quite likely that these are different from the rest - perhaps co-ops in other districts have different problems; perhaps those close to a big city are different from those in more remote areas. Or you could do a tour of the country by road, calling in on co-ops on the route. But again, these co-ops might not be typical - perhaps those close to main roads have fewer transport problems than those in inaccessible villages; perhaps it is only the larger, better organised ones that have someone in their offices when you call. Or you could write to 200 of the co-ops in the hope that about 50 would reply, arranging a time for you to visit. But, once more, these 50 would probably not represent the total 700 - those who replied would tend to be the more efficient ones, or they would be those in areas that had a good postal service.

Any of these methods is going to give a sample of 50 which have certain characteristics making them unrepresentative of the total 700. Each method favours certain co-ops (those near the capital, those near a road, those with efficient secretaries) at the expense of the others. To get a representative sample, you want a sampling method that does not favour any type of co-op over any other, i.e. a method in which every co-op has an equal chance of being included in the sample. This type of sample, in which each of the items being sampled (i.e. co-operative societies, in my example) has an equal chance of being selected, is known as a random sample.

As with many terms in statistics, there is some confusion between the technical use of the word 'random' and the use of the word in ordinary language. Suppose someone said, 'The menu was written in Chinese, which I couldn't understand, so I just chose a few dishes at random.' He would mean that he chose the dishes in a careless, haphazard way. But if a statistician says, 'I selected a random sample of 50 co-ops,' he does not mean that he chose them in a careless way. Quite the reverse. He means that he ensured that every one of the 700 co-ops had an equal chance of being selected and that, to achieve this, he used a random sampling procedure.

There is a big advantage that a random sample has over non-random samples (such as the ones I described at the beginning of this co-ops example). Say you visited a non-random sample and you found that 12 of the 50 co-ops (24%) had illiterate committee members. This would obviously have implications for any distance-teaching programme that you designed for them, but how reliable is this figure of '24%'? In other words, do 24% of all the co-ops have illiterate committee members, or is the real figure higher, or lower, than this? With a non-random sample, all you know is that the sample is probably not representative and therefore that '24%' is probably wrong, but you don't know how wrong it might be. Maybe there are only 12 co-ops like this in the whole country and you just happen to have got all of them in your sample. Or maybe you have tended to concentrate on co-ops with more educated people, and the proportion of co-ops with illiterate committee members in the whole country is very much higher than 24% - 50%, perhaps, or even 90%. You just don't know.

If your sample is random, you still know that '24%' might be wrong. But, because the sample is random, you can calculate how wrong the '24%' might be. In fact, in this example, you could be confident that the true figure for all 700 co-ops was between 12% and 36%; you could be fairly sure that between 12% and 36% of all co-operative societies in the country had illiterate committee members. I will say more about this in Chapter 11 (under 'confidence limits') and explain in Appendix 1 how to do the calculation. For the moment, the important point is that, with a random sample, you have a way of estimating how reliable the results are; with a non-random sample, you generally don't.

Simple random samples and systematic random samples

How do you draw a random sample of 50 co-ops out of a total of 700? The best way is to use a table of random numbers. You will find a page of random numbers in Appendix 7, and you will find more, if you need them, in a book of statistical tables (see Appendix 5) which you can buy, or borrow from a library. The table will contain strings of numbers, like this:

1 3 2 2 5 7 8 6 6 7 4 7 0 0 7 2 8 5 8 0 2 1 7 3 5

It might seem absurd to consult a book to get a list of random numbers. You might think you could produce your own list of random numbers by just writing down any numbers that came into your head. But, in fact, a list that you produced like this would not be completely random. (If you don't believe this, write your own list of 500 numbers, and then copy a list of 500 numbers from a table of random numbers, and take them to a statistician; he should be able to tell you which is which.)

You have to devise some way of using the random numbers to select respondents. In the co-ops example, you could give each co-operative society a three-digit code number, from 001 to 700. You would then take any line of the random numbers in groups of three. Say you were using the line that I have used in the illustration in the last paragraph. The first three digits are 132, so you would select co-op number 132 as one of your sample. The next three digits are 257, so you select co-op number 257 as the next one. (If you got a number between 701 and 999, or 000, you would just reject it, and take the next group of three digits.) You would carry on taking the random numbers in groups of three until you had selected 50 co-ops. These 50 would be a random sample - technically, a 'simple random sample'. (If you use the same table of random numbers on different occasions, start at a different point in the table each time. If you always began from, say, the top left of the table, you'd be using the same sequence of numbers each time, and you don't want to do that.)

It is theoretically possible that your random sample of 50 would be completely unrepresentative. For example, it is possible (theoretically) that all 50 would be in the same district. But this is extremely unlikely - so unlikely that you need not worry about it.

There is a variation on this method known as a 'systematic random sample'. You begin by dividing the total by the sample size; in the co-ops example, this is 700÷50=14. You choose a number at random between 1 and 14. This is the first co-op in your sample, and then you choose every fourteenth one in the list after that first one. The proportion in your sample - in this case 1/14 - is called the 'sampling fraction'.

When using this method you must make sure that there is no pattern in the list with which your sampling fraction coincides. Suppose the co-ops in your list were organized into groups of seven, for some reason, such that every seventh was a handicrafts co-op. By taking every fourteenth you might end up with a sample composed entirely of handicrafts co-ops, which would not be representative of the total. As another example, suppose you picked every tenth name from a list of schoolchildren. If the list happened to be arranged in the order boy-girl-boy-girl, you would get a sample composed entirely of boys or entirely of girls, depending on your starting point.

Stratified samples

Suppose you were going to interview 100 headteachers to find out what use they made of educational radio programmes in their schools. Let's say there were 2 000 schools in the country - 1 800 primary and 200 secondary. A representative sample of 100 schools would contain 90 primary and 10 secondary, i.e. the same proportions as in the total 2 000. If you used the first of the random sampling methods I described, to get a simple random sample, you would probably find that your sample contained approximately the right proportions, but not exactly the right proportions - you might end up with 8 or 11 secondary schools in the sample rather than exactly 10.

A slight improvement on this would be to divide up the original list of 2 000 schools into primary and secondary. Then you would take a random sample of 90 primary schools out of the total 1 800 primary schools, and a random sample of 10 secondary schools out of the total 200 secondary. In this way, your final sample of 100 schools is still random, because you have used a random method to select the schools, but it contains exactly the right proportions of primary and secondary. When you divide your original list into different groups and take a sample from each group separately, the result is called a 'stratified sample'. (The word 'stratum' - plural 'strata' - literally means 'layer'. You could say, for example, that the earth's crust was formed of strata of different types of rock. The idea is that, in my example, the primary schools form one stratum of schools, and secondary schools form another.)

It is possible to take more than one thing into account when stratifying a sample. Suppose your list of schools also showed how many pupils each school had, and you calculated that 940 of the schools -900 primary and 40 secondary - had fewer than 100 pupils. A table of the 2 000 schools would look like this:

	Number of schools	%
Primary, with 100 or more pupils	900	45
Primary, with under 100 pupils	900	45
Secondary, with 100 or more pupils	160	8
Secondary, with under 100 pupils	40	2
Total	2 000	100

You could divide the list of schools into these four groups and take a random sample of one-in-twenty from each group separately. The resulting sample would contain 45 larger primary, 45 smaller primary, 8 larger secondary and 2 smaller secondary, i. e. exactly the right proportion of each of the four groups. You would describe this sample as being stratified by type of school and size of school.

Stratifying would be particularly useful for drawing samples of the population in a country where people fell into distinct groups. For example, if the country was divided into several regions, and the inhabitants of each region had their own language and way of life, you could stratify your sample so as to include the correct proportions of each group.

Stratifying makes a sample more representative. You make sure that the sample matches the total exactly in one or two important respects; this means that it will probably match the total more closely in other respects also.

Multi-stage sampling (cluster samples)

In order to use a random sampling method such as the ones I have described for choosing the co-ops or the schools, you have to have a list from which to make your selection. In the co-ops example, you could select 50 from a list of all 700 co-ops in the country. However, it often happens that there is no such list.* If you wanted to interview a sample of rural housewives, for example, there would not be any list of all the rural housewives in the country. So how do you select a sample?

In this case, you might obtain a list of all the villages in the country. You could draw a sample of villages using the methods I have described, and then you would visit each selected village and interview a number of housewives in each one. This is drawing a sample in two stages - first a sample of villages, then a sample of housewives in each selected village.

---

* You would have the same problem, in effect, if you had reason to think that your list was seriously inadequate. Suppose your list of co-ops is ten years old and that, since the list was drawn up, 300 of the co-ops on it have ceased to exist and 400 new ones have been created. A random sample of 50, drawn from this old list, would contain some non-existent co-ops and, more seriously, would omit all the newer ones.

As another example, suppose you wanted to interview a sample of schoolteachers, but there was no list of all the schoolteachers in the country. If you could obtain a list of all the schools in the country, you could draw a sample of schools, and then interview a sample of the teachers in each selected school.

A word of warning about the second stage. Say you send interviewers to a village to interview 20 housewives. If it is a small village which has only 20 housewives in it, there is no problem; the interviewers interview all the housewives. But if there were 60 housewives in the village, the interviewers would have to choose which ones to interview. Do not leave it up to the interviewers to choose whoever they want. If it is left to them, the interviewers will tend to choose people who are more articulate or more approachable, and they will avoid people who seem to be aggressive or very shy. In other words, their sample will not be representative. You have to give the interviewers some clear instructions in advance, such as, 'Each interviewer is to take a different approach road into the village, visiting every second household along the road.' (For a more detailed example of this, see Chapter 6.)

Samples of this type are called 'cluster samples'. You take advantage of the fact that people are organised, or clustered, into larger units - villages, schools or whatever. You first draw a sample of these larger units, or clusters, and then a sample of people within each cluster.

Heads of households in rural Lesotho (an example of sampling)

Different countries will present different sampling problems, of course, but I will illustrate the points I have made by describing how we might take a sample of rural heads of households for a survey in Lesotho. (I will simplify the details, but only slightly.) If you don't want this amount of detail, skip this section.

We would have to use two-stage sampling, for two reasons. The first is that there is no reliable and up-to-date list of all heads of households that is available to the public. The second is that Lesotho is very mountainous. If we selected a sample of 240 heads of households scattered in 240 different places all over the country, it would take years to visit them all. In practice, we have to select a small number of villages (generally between six and ten) and then interview a number of people in each village. (It can take several months to visit just ten villages.)

One volume of the Lesotho census report contains a list of all the villages in the country. (There are several thousand.) The list gives a four-figure code number to each village, indicates where the village is and shows what its population is. A part of the list would look something like this:

Page 47. District 8 : Metsoalleng

	Village number	Name	Population
Sub-district 840	8401	Ha Mofokeng	307
	8402	Ha Makoae	253
	8403	Ha Molelle	187
	8404	Ha Morolong	166
	8405	Ha Maime	142
Sub-district 841	8411	Ha Nailane	282
	8412	Ha Tilo	133

The two items of information provided for each village - its whereabouts and its size - could be used to produce a stratified sample of villages. Let's say we had the time and money to visit eight villages. We would stratify, first, by district. One can calculate from the list that about three-quarters of the population live in the lowland districts and about one-quarter in the mountain districts. If we took six of our eight villages from the lowlands and two from the mountains and then interviewed the same number of people in each we would have the correct proportions of mountain people and lowlands people in the final sample.

Taking the village lists for the mountain districts first, we would calculate how the population was divided as regards size of village. We might find, say, that half the mountain population lived in villages of fewer than 200 people, and half in villages of 200 or more. So we decide to take one mountain village with a population under 200, and one with a population of 200 or more. We make the actual selection using a table of random numbers. We would go through the random numbers, taking them in groups of four, looking up each four-figure number in the village list. This might go as follows:

7955.	The village with this code number is Ha Tsekoa. Mountain village. Population 183. We take this as our one small mountain village.
9972.	There is no village with this code number.
7946.	Village Ha Lesenya. Mountains. Pop. 196. We already have our one small mountain village, so we reject this one.
8327.	Village Ha Rakhoba. Mountains. Pop. 327. We take this as our one large mountain village.

Next we turn to the lowlands districts. We want six lowlands villages in the sample, so we calculate how the lowland population is divided into sixths according to village size. We might get the following results:

One-sixth of the lowlands population lives in villages of 150 people or fewer
" " " " 151-210
" " " " 211-250
" " " " 251-320
" " " " 321-460
" " " " 461 or more

Using the table of random numbers as before, we select one lowland village in each of these size-groups. Along with the two mountain villages, this gives us our sample of eight villages.

The second stage of the two-stage sampling is done by the interviewers in the villages. We instruct them to visit the same number of households - say 30 - in each village. The precise instructions would vary slightly from one village to another, depending on the size of the village. In Lesotho, the average size of a household is about five people, so you would expect to find about 30 households in a village with a population of 150 people. In villages of about 150 people, we would instruct the interviewers to visit every household; in villages of about 300 people, they should visit every second household; in villages of about 450 people, every third household, and so on.

The result of all this would be a random sample of 240 rural heads of households, divided in the correct proportions between mountains and lowlands, and between villages of different sizes.

It might seem that eight villages are not enough when the purpose of the survey is to provide information about the entire rural population. The number of villages you need depends on how much the villages differ from each other. In Lesotho, a sample drawn from just eight villages would be adequate for most purposes. This is because Lesotho's rural population is remarkably homogeneous; that is to say, people in one part of the country do not differ very much from people in another part of the country. In several surveys on different topics, we found a surprising consistency in results from villages scattered around the country. If you were doing a survey in a country where people differed markedly from one part of the country to another, you would have to stratify the sample so as to get the right number of interviews from each part, and you would have to take an adequate number of villages within each part.

How many interviews?

I have not yet said how many people you ought to interview. It depends on what you want to get from the survey. If you were pre-testing a pamphlet, a sample of 15 or 20 people could be adequate; at least they would be enough to show if there was something seriously wrong with the pamphlet. For other purposes, you might want a sample of 200 or 300. It is unlikely that you would ever want a sample much larger than that, for the types of research I described in Chapter 1. It will be easier to describe how you decide on the number of villages or the number of interviews after I have explained some statistics, so I come back to these questions in Appendix 1.

It is more important to have a good sample than to have a large one. Magazines often contain questionnaires (about violence on television, for example); the readers are invited to fill in the questionnaire and send it back to the magazine publisher. The publisher counts the replies and publishes the results in the next issue of the magazine. If the magazine has a big circulation (over a million readers, say), the publisher may receive several thousand questionnaires. But the results (e.g. '95% of people think that there is too much violence on television') are almost worthless, because the sample is not representative. The questionnaire will have been completed only by readers who had the time, or who felt strongly about the questions, or who enjoy filling in questionnaires. The results from a properly selected sample of only a hundred people would almost certainly be more reliable.

Two last pieces of advice on sampling. Some sampling problems are quite difficult. If you have a problem, consult a statistician. There will be statisticians in the university or in the government's department of statistics. (But be on your guard, see my remarks on 'Consulting documents and experts' in Chapter 2.)

Many pieces of research have to be done quickly and with limited resources. Perhaps you do not have the time or the transport to send interviewers to remote villages on the other side of the country. In such cases, get as good a sample as you can. If you are pre-testing some material that is intended for farmers, then at least take it out to some farmers; don't pre-test it on schoolchildren. If the interviewers can only go out for one day, then get them to visit at least two different places and preferably three. And instruct the interviewers to sample in a systematic way - every second household, or whatever; don't leave it to them to interview anyone they want.

5 Questionnaires

This chapter describes how to write a questionnaire for a social survey. Most of the chapter is about questionnaires that are designed to be used by interviewers but the same principles apply to questionnaires that are filled in by the respondents themselves.

The first page *Survey title, the organisation's address, the serial number, the interviewer's preamble.*

Break down general questions into specific ones *Specify as precisely as possible the information you want.*

Closed and open questions *Closed questions restrict the range of possible answers. Open questions allow many possible answers.*

Categorising answers in advance *Writing the possible answers on the questionnaire makes the interviewing and analysis easier, but it requires some care when writing the questionnaire.*

Pre-coding *Putting code numbers on the questionnaire makes it easier to process the questionnaire later.*

Bad questions *Avoid vague questions and leading questions. Be careful with sensitive topics and with 'forced-choice' questions.*

Branching *Not all the questions apply to all the respondents. Branching instructions tell the interviewer which questions to ask.*

The last page *Information on age and sex. Interviewer's winding up.*

Self-completed questionnaires *Further points for when the respondents are to complete the questionnaire themselves.*

Translation *Sometimes a questionnaire is written in one language and then translated into another before being used by the interviewers. Some care is needed to ensure that the translation is as accurate as possible.*

When you process the results of a social survey, you add together the answers given by different people to obtain results for the whole sample in quantitative form. Say you interview 100 sheep-farmers and you ask them, 'Have you had any of your sheep dipped in the last twelve months?'; you count up their answers and you find, say, that 28 said Yes and 72 said No. In order that you can add together the answers of different farmers to arrive at this result, it is essential that all the 100 farmers are asked exactly the same question. If some farmers were asked, 'Have you had any of your sheep dipped in the last twelve months?' and others were asked, 'Have you had any of your sheep vaccinated in the last twelve months?' it wouldn't make sense to add their answers together.

In this respect, a survey interview is quite different from the sort of discussion I was describing in Chapter 3. Instead of inviting each farmer to tell you his opinions - whatever they may be - about sheep-farming, you ask each farmer a certain set of questions that have been prepared in advance. To make sure that each person is asked exactly the same questions, you write the questions in the form of a questionnaire. Generally you print many copies of the questionnaire and the interviewer uses a fresh copy for each person he interviews, reading the questions from the questionnaire and also marking the answers on it.

So far, I have been describing the sort of survey in which the people answering the questions (who are called 'respondents') are interviewed by interviewers; the interviewer has the questionnaire, reads out the questions to the respondent and marks the respondent's answers on the questionnaire. You can also conduct a survey by giving each respondent a questionnaire and asking him to fill it in himself; these are called 'self-completed questionnaires'. The main principles apply to both types of questionnaire, but there are some differences. For most of this chapter, I will describe interview-questionnaires, and then I will say a little more about self-completed ones towards the end.

The first page

It is useful to put the title of the survey at the top of the first page. A research division might do ten or twenty surveys in a year; you need to be able to see at a glance which survey a questionnaire belongs to. Also write 'Please return to .....' (then give the address of the organisation). If the interviewers lose some questionnaires while doing the fieldwork, someone might find them and send them to you. At the top right-hand corner, leave a space for the serial number. I will explain the purpose of this when I talk about processing the results (see the section on 'Coding onto transfer cards' in Chapter 7.) Include places for the interviewer to write his own name and the date of the interview. These may be useful later when checking on the interviewing.

In general, you do not need the name of the respondent. In Lesotho, we have found that it is better not to ask for people's names. Like many other newly independent countries, Lesotho has been troubled by political conflict. People are reluctant to give their names to strangers, especially strangers who look official. Our interviewers have always reported that people were much happier to be interviewed when they saw that the interviewer was not writing their name and in fact did not even want to know their name.

Underneath all these items, you write the interviewer's preamble. This is what the interviewer says to each respondent before he asks the first question. It should include the name of the organisation and perhaps a brief description of it. It should mention the authorisation for the survey - that you are acting with the permission of the government, for example, or of the local chief. It should explain briefly the purpose of the survey. It should reassure the respondent that his answers will not be discussed with other people. Finally, it should include an invitation to the respondent to ask any questions before the interview begins. It might also include other points, of course, depending on the topic of the survey.

As an illustration, the first part of a questionnaire by LDTC on poultry-keeping might look like this (the final version would be in the Sesotho language, of course):

Please return to:
Lesotho Distance Teaching Centre
PO Box 781
MASERU

Serial number

POULTRY-KEEPING

Interviewer's name ....
Date of interview ....

My name is .... I am working for the Lesotho Distance Teaching Centre. This is an organisation which uses booklets and radio programmes to teach people. For example, we have produced booklets on cookery and farming. We are thinking of writing a booklet about keeping poultry, but, before we do, we want to find out people's opinions about poultry keeping. Our work is approved by the Ministry of Education, and we have obtained the permission of chief .... to visit people in this village. I will note down your answers on this questionnaire, but I will not write down your name. I will not discuss your answers with anyone else in the village. Are there any questions you would like to ask me before we begin?

Break down general questions into specific ones

Next comes the hard part. You have to decide exactly what questions you want to ask. In Chapter 2 I said how important it is to specify at the outset what you want the research to find out. If you have not done that, you will start running into trouble when you try to write the questionnaire, because you will not know what questions you want to ask.

Let's assume that you have a fairly clear idea of who you are going to interview and of what information, in general terms, you want to get from them. For example, you have decided to interview heads of households and you want to ask each one whether he has any poultry and, if so, how he keeps them. Or you have decided to interview married men and women aged between 20 and 45, and you want to find out if they have heard of contraception, what they think about it and whether they have used a contraceptive method. You have to break down these general questions into more specific ones. It is no good to ask, 'How do you keep your poultry?' One man might tell you about how he feeds them, another about where he shelters them and another about how he cures their diseases. The result will be a jumble of information. You have to decide precisely what you want to know.

It can be helpful to do this in stages. First, jot down the various topics that you want to cover in the questionnaire. For the poultry survey you might write:

How many birds?
Food? Shelter? Diseases?
Traditional or improved breed?
Kept for eggs or meat?
How many eggs?
How often slaughtered?
Home consumption or for sale?
Use of Ministry of Agriculture
poultry experts?

Then you can take each of these topics and jot down more detailed points about each one. Under 'Food?', for example, you might write:

Give food to birds or do birds fend
for themselves?
Household scraps or food specially
made up?
Buy special food for birds?
- Where?
- What price?

With a list of points like this, you can begin to write the questionnaire. This means turning each of these points into a clear, simple, answerable question.

Another way to approach the writing is to think of the different sorts of people who will be answering the questions. In the family planning example, you might think of different questions that you want to put to the men and to the women, or to the younger couples and to the older couples, or to those with children and to those without. As another example, suppose you are designing a survey of housewives, to evaluate a cookery book that has been on sale for some time. You might begin by jotting down the following:

Some won't have seen the book.
Some will have seen it, but not
bought or borrowed it.
Some will have got a copy but
not read it.
Some will have read it but not
used it.

One can then begin to think of questions appropriate to each group.

Closed and open questions

I said just now that you should write questions in the questionnaire which are clear, simple and answerable. That is easier said than done. You need to give a lot of thought to the precise wording of the questions. Do not expect to write a questionnaire in half an hour. It can take several days to produce a good first draft, and then several more to improve it, though half a day might be enough for a short and simple one.

There are different types of question, each with its pros and cons. One distinction is between questions which restrict the range of answers ('closed questions') and those which allow many different answers ('open questions'). The following are examples of closed questions:

Do you have any poultry?

Have you visited a doctor in the last month?

Would you be interested in a booklet on bicycle maintenance?

In the cookbook, were there too many recipes, not enough recipes, or about the right number of recipes?

Nearly all the answers to these questions will be straightforward - 'Yes' or 'No' or 'Not enough recipes', or whatever. There might be a few answers that are not simple, such as, 'I don't own any poultry myself, but I am keeping a few birds for someone else.' But you would expect the great majority of the answers to fall into one or two categories. The following are examples of open questions:

What food do you give to your birds?

Who would you go to for advice?

What topics would you be interested in for future booklets?

Why have you not tried any of the recipes?

You might get many different answers to each of these questions.

In general, I prefer closed questions. You decide exactly what information you want and you ask for it. For example, if you are particularly interested in knowing what proportion of poultry owners buy special poultry food for their birds, it is better to ask a specific question about it, e.g. 'Have you bought special poultry food for your birds in the last six months?' If you ask only an open question, such as 'What food do you give to your birds?', you might not discover what proportion bought special food. An owner who buys special food only occasionally might say 'Household scraps' in answer to the open question, meaning that he generally feeds his birds on household scraps. I would prefer to break down the open question into a series of closed questions, such as:

Do you give your birds any food, or do they fend for themselves?

Do you give them household scraps?

Do you give them special food bought from a shop?

You can then ask an open question at the end of this series, such as

What sort of food do you give them most often?

There are times, however, when an open question is preferable. For example, if you were pre-testing a drawing of a road accident, which was to be used in a road safety poster, it would be foolish to ask, 'Do you think this is a picture of a road accident?' The whole point of the pre-test is to find out what people think the picture represents, without any prompting. So you would ask, 'What do you think this picture shows?'

Categorising answers in advance

There are also different ways of recording people's answers. You can write down the different possible answers on the questionnaire, so that the interviewer just ticks the one that the respondent gives. Or you can leave a space for the interviewer to write in the respondent's answer. The following are examples of the first kind:

Did you listen to the radio at 7.00 p.m. last Thursday?

Yes ☐ No ☐ Can't remember ☐

What is your age?

20 or under ☐

21 to 30 ☐

31 to 40 ☐

41 or over ☐

The following are examples of the second kind:

What punishment do you think should be given to drunken drivers? ....

What is your age? ....

The first kind are easier for the interviewer. They are also easier to handle at the later stage of counting the results. But you must be careful when writing them. First, it is important to cover the full range of answers that the respondents might give. In the following example there are not enough categories:

What do you think is the ideal number of children for a couple to have today?

Three or fewer ☐

Four or more ☐

There ought to be categories for people who say 'Don't know' or 'God will decide' or 'It's impossible to say'. If you do not provide enough categories, the interviewers will try to squeeze people's answers into categories to which the answers do not really belong. If you are not sure what answers people might give, include categories such as 'Other' and 'Don't know'. If you were interested to know what these 'other' answers actually were, you could ask the interviewer to write them in, like this:

Other ☐ (Explain ....)

Secondly, make sure that the categories will give you all the information you want. For example, you might divide peoples' ages as follows:

20 or under ☐

21 to 40 ☐

41 or over ☐

At the stage of analysing the results, are you going to want to look at people aged 61 or over as a special group? If so, you must include that category on the questionnaire. If you leave it as in the example, the 61-or-overs will be all mixed up with the 41-to-60 people and you will not be able to look at them separately.

Questions which do not have the answers written out in advance are more tedious for the interviewers. This is a serious problem if each interviewer has more than about ten interviews to do. Take a question like this, 'What do you think are the advantages of child-spacing?' For the first few interviews, the interviewer will faithfully record the respondent's answers, as, for example, 'It gives the mother time to feed her child well and care for him so that he will grow strong and healthy,' or 'Prolonged breastfeeding makes a healthy child.' By the fiftieth interview he will be summarising all replies of this type in a short phrase such as 'Health of child'. The interviewer tends to ignore what he considers to be unimportant differences in the words that people use and jot down what he thinks is the main point; in effect, he devises his own set of categories. Unfortunately, different interviewers will devise slightly different sets of categories, so they will appear to get different results. It is better that they should all use the same categories, and you can ensure this by writing them on the questionnaire.

It must be admitted that putting people's answers into categories can be, in a sense, unfair. People's answers to a question may vary in many subtle ways, and this variety is lost if you put these answers into just two or three categories. Regrettably, this is unavoidable. A report of a survey of 250 people does not look like this:

Respondent 001 said ......
Respondent 002 said ......
Respondent 003 said ......
and so on.

Such a report would be unreadable. The report has to contain statements like, '42% of the respondents said this .....' You have to lump people's answers together in some way in order to make them manageable.

At some stage, you have to invent categories for the answers to a question and to put each person's answer into one of these categories. If you don't do this when you write the questionnaire, you have to do it later. If you leave a space for the interviewer to write down each person's answer, you have to go through all the questionnaires later on, reading each person's answer and putting it into a category. This can be a lengthy and tedious job.

In general, especially for surveys of more than about 40 people, I would recommend using closed questions and writing out the answers in advance. This requires more care and thought at the stage of writing the questionnaire, but the results are likely to be more manageable and more reliable.

Pre-coding

At this point, I have to mention something that I will explain in more detail in Chapter 7. If you intend to use computer cards at the later stage of processing the results, you have to convert people's answers into code numbers. In fact, it is often convenient to do this even if you don't use computer cards. Take the following question as an example:

> Were there too many recipes in the book, too few recipes, or the right number of recipes?
>
> Too many ☐
> Too few ☐
> Right number ☐
> Don't know/No opinion ☐

In order to put people's answers to this question onto computer cards, you would have to give a code number to each of the answers, like this:

> Too many - Code 1
> Too few - Code 2
> Right number - Code 3
> Don't know/No opinion - Code 4

You can save time by putting these code numbers on the questionnaire. So you would write out the question like this:

> Were there too many recipes in the book, too few recipes, or the right number of recipes?
>
> Too many 1
> Too few 2
> Right number 3
> Don't know/No opinion 4

The interviewer rings the code number; for example, if the respondent answers 'Too few recipes', the interviewer rings the number 2. Questions with the answers set out like this are called 'pre-coded' questions.

Bad questions

There is no simple technique for writing good questions. The best I can do is warn you about various types of bad question.

Questions can be too vague. You may want to find out people's attitudes to family planning, but it is no good to ask people, 'What is your attitude to family planning?' Some people might think this question is about their personal experience of using a contraceptive method; others might think it is about their moral or religious principles; others will be people who have never even heard of family planning.

Similarly, a phrase in a question can be too vague. For example, 'Have you listened to an educational radio programme in the last week?' is not clear because different people would have different definitions of 'educational radio programme'. Some might think it means a programme for schools; others might include programmes for adults on nutrition or farming; others might include any programme that gives information, such as the news.

It is often useful to specify a time period in a question. If you were broadcasting radio programmes to help correspondence students and you wanted to know how many were listening, it would not be much use to ask 'Have you listened to our radio programmes?' A student who had listened to only one programme ten months ago could answer Yes. It would be more useful to ask 'Have you listened to any of our radio programmes in the last month?' or 'Did you listen to the mathematics programme broadcast on Tuesday evening last week?'

Questions should be answerable. Do not ask the respondent to perform feats of memory. For example, the question, 'How many letters have you received in the last year?' could be impossible for many people to answer. Do not assume that respondents can make accurate assessments of distance or time in standard units; many rural people would not be able to answer the question, 'How far from your home are your fields?' And do not assume that they have knowledge when they may not. For example, 'What type of non-formal education facilities would you like to have in your village?' assumes that the respondents have some idea of what 'non-formal education facilities' might be; perhaps they have no idea at all.

People are unlikely to give their genuine opinion if the question has an obvious 'right' answer, or a polite answer, such as 'Are you in favour of controlling soil erosion?' or 'Have you found the course interesting?' Nearly everyone will say Yes, even though they do not care about soil erosion at all, or they have found the course extremely boring. In particular, the wording of the question should not guide people to a particular answer, like this, 'The population is increasing all the time, but there is only a limited amount of agricultural land in the country; do you think this is a problem?' The question clearly expects the respondents to answer Yes, and most of them will.

Naturally, people tend to give the answer they think you want, or the answer which will make a good impression. A nice example of this is provided by some questions which LDTC put to its own correspondence students. Those students who had a radio which received Radio Lesotho reasonably well were asked whether they had ever listened to LDTC's weekly radio programme and, if so, whether they had done so in the last month. About 80% said they had listened to LDTC's programme at some time and about 50% said they listened regularly - three times or more in the last month. The programme seemed to have a regular audience. However, the students were then asked to say at what time the programme was broadcast and they were given four possible answers to choose from, e.g. '7.15 p.m., Monday' '6.30 p.m. Wednesday' etc; the proportion who knew the correct answer was only about 10%, which suggested that they were not such keen listeners after all.

Sometimes there is a particular reason, connected with the topic, why people may not give a straight answer to a question. For example, if people who own radios are supposed to buy radio licences, they may not give a truthful answer to the question, 'Do you own a radio?' Or if a man's tax is based on the number of cattle he owns, he may not give the right answer to, 'How many cattle do you own?' In Lesotho, it would be difficult to ask adolescent boys about their experience of initiation ceremonies, since these ceremonies are traditionally kept secret. At the same time, one need not be too shy of asking about delicate matters. If the interviewer wins the confidence of respondents, they will speak frankly about many aspects of their lives which they would not normally discuss with strangers. If you have to include questions on delicate topics, put them in the second half of the questionnaire, so that the interviewer and respondent will be more relaxed when they come to them.

If you think it is likely that people will not give a straight answer to a question, put the question in a different way or ask a slightly different question. Instead of 'Do you own a radio?' you might ask, 'Have you listened to the radio in the last week?' Rich cattle men who would refuse to answer the question, 'How many cattle do you own?' might answer the question, 'Do you own more than ten cattle?' (Perhaps this information would be sufficient for you.) Instead of 'Have you found this course interesting?' you could try 'List three good things and three bad things about the course.' If you were afraid that people would all say Yes to the question, 'Did you listen to our radio programme on mathematics last night?' you could say, 'Please tell me which of the following programmes you listened to last night: Hygiene Hints at 6.30, News at 7.00, Mathematics at 7.30, Record Requests at 8.00.'

Another type of question that you should avoid, or at least use with caution, is one that forces the respondent to make a choice. For example, if you asked, 'Are you for or against population control?' the results might suggest that people have strong views on the subject ('90% of people are against population control'), whereas the truth might be that most people have very little idea of what 'population control' means and have no strong views for or against. Questions of this kind can be useful, so long as you interpret the results carefully. For example, we asked housewives to choose between various possible topics for future booklets, with questions such as, 'Would you be more interested in a book on thatching, sheep-shearing or bicycle maintenance?' If the majority voted for sheep-shearing, that would show that a booklet on sheep-shearing was more likely to be popular than booklets on the other two topics, but it would not guarantee that a sheep-shearing booklet would be a best-seller, or that the others would be unpopular.

Finally, if you want a statistic (e.g. 'What proportion of farmers use fertiliser?'), collect the results which will enable you to calculate the statistic; don't collect opinions about what the statistic is. In this example, you should visit a representative sample of farmers and ask each one whether he uses fertiliser. Then you can count up the proportion of the sample who use fertiliser. You should not ask a lot of farmers the question, 'What proportion of farmers use fertiliser?' A farmer will know a great deal about his own farm, but he does not necessarily have statistical information about farmers in general. A mistake of this type was made by a correspondence college which wanted to assess the need for correspondence education in the country. What they could have done was to interview potential correspondence students (such as secondary school dropouts, unqualified teachers, junior civil servants) to assess their needs. Instead, they interviewed a number of public figures (employers, headmasters and so on) about what they thought the needs were. Finding out opinions about needs might be useful, but it is not the same as finding out about the actual needs.

Branching

With some questionnaires, all the respondents answer all the questions. If you had run a three-day course on book-keeping for treasurers of co-operative societies, you might administer a questionnaire to all the participants at the end containing some test questions on book-keeping and some other questions on the usefulness of the course, the quality of the accommodation and so on. All the questions would apply equally to all the participants.

However, it often happens that parts of a questionnaire apply to some of the respondents but not to the others. Take the following example:

1. Did you listen to the radio programme on mathematics yesterday evening?

   Yes ☐
   No ☐
   Can't remember ☐

2. Was the quality of reception good, middling or poor?

   Good ☐
   Middling ☐
   Poor ☐
   Couldn't say/Can't remember ☐

It would obviously be silly to ask question 2 if the respondent has just told you that he did not listen to the programme. Or take the following example, from a questionnaire on family planning for married men and women aged 20 to 45:

7. Do you have any children?

   Yes ☐
   No ☐

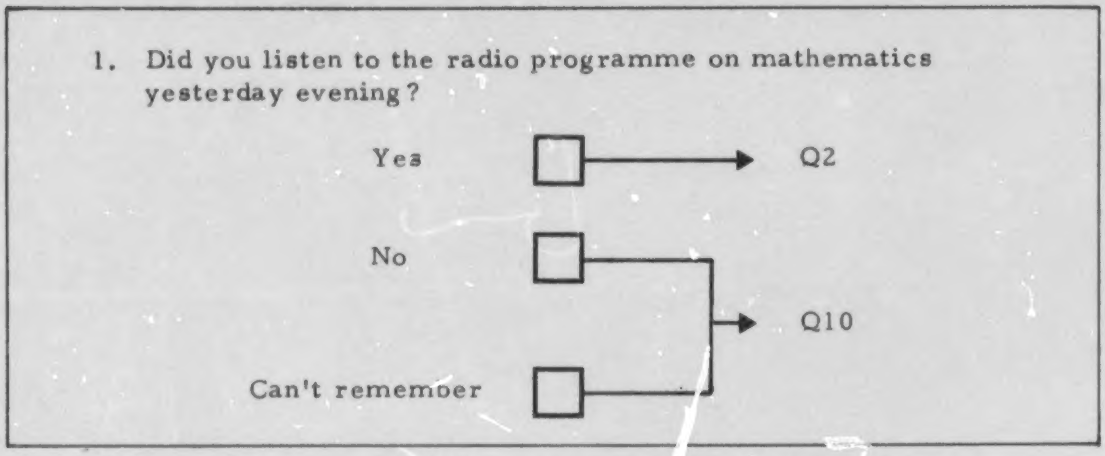
8. In what year was your last pregnancy?

   ....

Question 7 could be put to both the men and the women, but question 8 could apply only to women.

You want the interviewer to skip certain questions if they are inappropriate. You cannot expect an interviewer to decide on the spot which questions are appropriate and which are not; you have to give him some instructions on the questionnaire.

You could write instructions in this form, 'If No or Don't know, skip to question 10.' In my experience this is not satisfactory, as it takes the interviewer a little time to read this instruction and then to decide what he has to do, so it halts the flow of the interview. Interviewers find it easier if the instructions are written like this:



1. Did you listen to the radio programme on mathematics yesterday evening?

Yes ☐ → Q2

No ☐ → Q10

Can't remember ☐ → Q10

Instructions of this kind are called 'branching instructions' (or 'filters'). You have to think carefully about them. If you leave one out where you should have put one in, the questions may be embarrassing, or annoying or just silly, like asking a man about his pregnancies, or an illiterate about his reading habits, or a barren woman about her children. On the other hand, you don't want to skip too many questions by mistake. It can happen, with a complicated questionnaire, that you intended a particular question to be put to all the respondents and then you discover that, because of a branching instruction you put earlier in the questionnaire, some of the respondents have skipped that question. It is very annoying to discover this after the fieldwork has been done. In particular, if the draft questionnaire is revised several times in the light of comments from one's colleagues, with questions being added and deleted, you must check the branching instructions on the final version very carefully to see that the question numbers are correct.

The last page

If you have not already noted this information on the questionnaire, it is often useful to record the respondent's sex, age and education. You might arrange this as follows:

Now I'd like to ask you a few questions about yourself.

43. Interviewer note respondent's sex
Male ☐
Female ☐

44. What is your age?
21-30 ☐
31-40 ☐
41-50 ☐
51-60 ☐
61 or over ☐

45. Did you attend school?
Yes ☐ → Question 46
No ☐ → Question 47

46. At what standard did you leave school?
Primary 1 to 3 ☐
Primary 4 to 7 ☐
Post-primary ☐

Of course you will select the most appropriate categories for age and education, depending on the survey.

There are two reasons for collecting this basic information about each respondent. One is that it helps you to see how representative the sample is. The other is that you often want to compare men with women, older with younger people, and more educated with less educated people. I will explain this more in Chapter 9.

At the end of the interview, the interviewer should give the respondent the opportunity to ask him any more questions about the survey. And, of course, the interviewer should thank the respondent. You can remind the interviewer to do these things by writing them on the questionnaire:

Are there any questions that you would like to put to me?

Thank you for your co-operation.

At the end of the questionnaire, you might ask the interviewer to note down further observations, such as the size of the respondent's home (if this information were of any use to you). You should leave a space for the interviewer to make any comments on the interview, such as, 'Very difficult interview because respondent was drunk,' or 'Respondent's mother came in half-way through and insisted on adding her opinions, which possibly influenced respondent's answers.'

Self-completed questionnaires

If the respondents are all able to read, it is possible to give them the questionnaire so that they can write in the answers themselves. If you ran a short residential course for correspondence students, you might give them a questionnaire like this as part of the evaluation. One great advantage of self-completed questionnaires is that you can send them through the post. To do a survey of headteachers, for example, you might send questionnaires to a sample of headteachers by post, enclosing a stamped addressed envelope and asking them to complete the questionnaire and return it to you.

The principles of writing a self-completed questionnaire are much the same as for an interview questionnaire. With self-completed questionnaires, it is even more important that the questions should be unambiguous. Branching is more difficult, and perhaps impossible, on a self-completed questionnaire. If the respondents have had a fairly high level of education, they will be able to follow instructions such as 'If your answer to Question 2 is Yes, go to Question 3; if it is No, go to Question 15.' But less educated respondents will have difficulty with such instructions.

Since the questionnaire is to be completed by the respondents themselves, who are perhaps not familiar with questionnaires, it has to be clear and simple. You need to include clear instructions before the first question about how the respondents should record their answers, along these lines:

> Read each question carefully and decide what your answer is. Then mark your answer by putting a tick in the appropriate box. For example, if you decided that your answer was No, you would record this as follows:
>
> Yes ☐
> No ☑

If you send the questionnaire by post, you should enclose a covering letter with it. This letter has to do the job that an interviewer does in an interview survey, of explaining the purpose of the survey, describing briefly the organisation that is conducting it, gaining the respondent's co-operation and so on (see the section 'The first page', at the beginning of this chapter). It should also stress the need for a quick reply; if people do not fill in the questionnaire within a few days of receiving it, they are likely to forget about it. For the same reason, the questionnaire should be well laid out and nicely printed, so that it looks inviting rather than intimidating.

As with interview questionnaires, I favour 'tick the box' questions rather than open questions, partly because this forces you to think more precisely about the information you want and partly because the results are easier to handle. However, respondents often like to explain their reasons for their answers, so you should include some open questions and allow space on the questionnaire for them to write what they want to write, even though you may not intend to use these answers in the analysis.

Translation

In many developing countries, more than one language is used. The business of government is conducted in the official language - often English, French or Spanish - but this may be little understood by large sections of the population. So it often happens that a questionnaire is drafted in one language, but the interviews are conducted in another.

Some researchers leave it to the interviewers to make the translation on the spot - the interviewer has a questionnaire in English, for example, but asks the questions in Sesotho. This makes things easier for the researcher but it can ruin the survey. Different interviewers will translate questions in different ways, and the same interviewer may translate a question differently on different occasions. In effect the respondents are getting different versions of the questionnaire. But the basic idea of a survey is that, so far as possible, all the respondents are asked exactly the same questions - it is only on that basis that you can add together their answers at the end. It follows that you should translate the questionnaire in advance into the language in which the interviews will be conducted so that the interviewers have that version in front of them when they are doing the interviews.*

People who have never done any translation sometimes imagine that it is a simple matter of substituting the words of one language for the words of another, a virtually automatic process like converting prices from one currency into another; you just give your questionnaire to a translator, and back it comes saying exactly what the original said, only in a different language. Worse still, they occasionally think that it does not matter if the original is obscure or over-technical since the translator will make it all clear. But translation is not so straightforward. I will illustrate this with some examples from LDTC, where questionnaires were usually written in English and then translated into Sesotho.

A translator can misunderstand the original. For example, a question in English in an LDTC questionnaire was, 'What do you do on Sunday which is different from other days?' meaning, 'What do you do on Sundays which is different from what you do on other days?' The translator misunderstood the question and put it into Sesotho as,'What do you do on Sunday, which is the Lord's day?' This mistake was not spotted and the question was put to the respondents (rural teenagers) in that form. Not surprisingly, a high proportion replied, 'I go to church.' (We do not know whether they really did attend church regularly, or whether they were prompted to give this answer by the wording of the question.)

---

* In countries where several languages are spoken, the questionnaire may have to be translated into more than one other language. In that case it is even more important to have the different versions written, checked, and brought into line with each other as far as possible, before the interviewing begins.

The translator sometimes imports his own knowledge or preconceptions into the translation. Another question in that questionnaire was, 'Do you take turns with a brother or sister at attending school?' The follow-up questions were, 'Do you attend on alternate days, alternate weeks, alternate terms, or in some other way?' and then, 'Why do you take turns at attending school?' The translator happened to know, or to believe, that all children who took turns attended on alternate days (rather than alternate weeks, etc.) and that the reason for taking turns was to tend the family's cattle on the days off school. He therefore translated the first question as, 'Do you attend school, taking alternate days with a brother or sister, in order to tend cattle?' The follow-up questions were therefore useless and we never found out whether any child who took turns did so on any system other than alternate days, or whether they did so for any reason other than tending cattle.

Many expressions have overtones or secondary meanings as well as their straightforward meaning, and a translator does not always take these into account. (In fact, he may not appreciate the overtones of the English expressions in the first place if English is his second language.) In America and Britain, it is customary to put 'Confidential' at the top of the first page of a questionnaire in capital letters. I suppose this is a reminder to the interviewers and other members of the research team that they must treat the information confidentially. In our first survey at LDTC, we printed the Sesotho word 'LEKUNUTU' ('confidential') on the first page. However, when a chief asked to have a look at the questionnaire, this word had the wrong effect. Whereas the word 'confidential' is a reassuring word, the word 'lekunutu' means 'secret' and suggests that something suspicious is going on.

Even with straightforward meanings, there is not a simple correspondence between one language and another. This obviously applies to technical expressions; no-one would expect 'boron-deficient soil type' or 'carburettor diaphragm' to go easily into Sesotho. But non-technical expressions can also cause trouble. The sentence, 'Germs are living creatures so small that they cannot be seen,' is very difficult to put into Sesotho. The problem is that any of the words for 'creature' really mean 'animal' and it is absurd in Sesotho to talk about an animal too small to be seen. Even the English words 'he' and 'she' do not mean exactly the same as their Sesotho translation. The personal pronoun in Sesotho does not distinguish between male and female; it is really equivalent to the cumbersome English expression 'he or she'.

It is quite unrealistic to hope that a translator will take an obscure questionnaire in one language and turn it into a clear one in another.

Only a translator with an exceptional grasp of both languages and of the subject in question could do that. It is more likely that the translation will depart in various ways - perhaps important ways - from the meaning of the original. What can you do about this?

First, if the writer of the questionnaire doesn't understand the other language, he should try to learn something of it. Even if he learns only the rudiments, he will be able to see if a question has been left out, or put in the wrong place, or if some quite extraneous ideas have been introduced. He will also appreciate some of the big differences in the grammar; perhaps the other language has no conditional tense, for example, or no precise way of saying 'bigger than' and 'biggest'.

Second, he should write the questionnaire in close collaboration with the translator, so that the translator understands the intention behind the questions. Failing that, he should go over his draft with the translator before handing it over for translation, explaining what he wants to get out of the survey.

Try to arrange that the person doing the translation is translating from his second language into his first language. This is because it is much more difficult to write in a second language than to read it. For example I can translate a passage of French into English reasonably well, but I would make mistakes if I translated a passage of English into French. In general, a translator should always translate into his own first language.

When the translator has produced a draft, someone else should go over it, with the original, to check it. More elaborately, you could have two people translate it independently and then get together to compare their translations, but this might take too much time.

Another possible procedure is to give the translation to someone else to translate back again into the original language; you then compare the original with the back-translation. Unfortunately, this is not as helpful as it might appear. None of the English people on the staff of LDTC had a fluent command of Sesotho; so the person doing the back-translation was always another Mosotho, translating from his own language into his second language. As I have just mentioned, it is more difficult to do a translation this way round, so a high proportion of the discrepancies between the original and the back-translation were introduced by the second translator and did not indicate defects in the first translation. However, if it is an important survey and you have the time, it might still be worth doing.

Finally, you should look through the results of the pilot survey, if you do a pilot survey (I'll say more about this in the next chapter). If you get some peculiar results in the pilot, you may find they are caused by faulty translation of the questionnaire.

6 Survey fieldwork

This chapter is about organising the interviewing.

Selecting interviewers *An interviewer's personality is more important than educational qualificiations. It is generally better if the interviewers have no particular personal interest in the topic of the survey.*

Training interviewers *You prepare detailed instructions for the interviewers on the sampling and interviewing and then train them thoroughly to follow these instructions.*

The pilot survey *A test run of the questionnaire before proceeding with the full survey.*

Administration *Payment, expenses, insurance, supervision and the like.*

When the sample has been drawn and the questionnaire has been written, the next step is to put the survey into operation. This chapter is about organising the interviewing, the stage of a survey usually known as the 'fieldwork'.

Researchers often give the fieldwork less attention than it deserves; they put great effort into the questionnaire and, later on, into the statistical analysis, but they seem to think that the actual collection of people's answers is a routine operation that will, somehow, run its course on its own. The reason for this neglect, I suspect, is that the organisation of fieldwork is less prestigious than other parts of the work. But it is surely obvious that the interviewing - the point at which the information is actually gathered from people - is crucial. No amount of sophisticated analysis at a later stage can compensate for poor fieldwork.

Selecting interviewers

Interviewers do not have to have high educational qualifications. At LDTC we found that people with three years of secondary education could do the job perfectly well. Personality is far more important. An interviewer must be able to gain the confidence of a respondent. In coming to a household and in conducting an interview, he must be courteous and patient. If the interviewers stay for some time in the same place - spending several days in the same village, for example - they must also conduct themselves in a sober and responsible manner when they are not interviewing. If the interviewers fail to gain people's confidence or if they antagonise people, respondents might refuse to be interviewed, or they might give false answers, or they might even complain to the authorities.

For certain surveys, it might be important to employ interviewers of a certain sex or age or marital status. Younger interviewers might be better for interviewing teenagers, for example , whereas it might be essential to have older, married interviewers for a survey about family planning. Depending on how the interviewers are to travel around, it might be important that they are able to drive a Landrover or to ride a horse.

Occasionally, it is useful to have interviewers who have some special interest in the survey. When pre-testing a picture, for example, you might include the artist among the interviewers. However, if you want unbiased information from the survey, it is usually better to avoid interviewers who have a special interest in the result. For example, a survey of farmers in one country used agricultural demonstrators as the interviewers. One of the questions was, 'What is your main source of agricultural information?'; according to the published results, almost 100% of the respondents said 'Agricultural demonstrators'. One does not know whether the respondents actually said this, out of politeness to the interviewers, or whether the interviewers recorded this answer regardless of what the respondent said, in order to convey a good impression of their activities. It was, almost certainly, a false result.

If it is possible, it is a good idea for the researcher to do some of the interviewing himself. He will experience any problems with the questionnaire and he will appreciate some of the subtleties in people's answers which are lost in the process of recording them in the questionnaire.

The number of interviewers you need will depend, of course, on the number of interviews you want to do and also on the way you divide up the interviewing among the interviewers. I have mentioned before that, for surveys in rural Lesotho, transport was a major problem. Consequently, we hired a very small number of interviewers - perhaps just three or four - and sent them around the country as a team, usually travelling in a Landrover. If transport was not such a problem, you could hire more interviewers and send them to different places in ones or twos. In a larger country, you might hire separate teams of interviewers in each region.

The disadvantage of having a small number of interviewers is that each one has to do a lot of interviews, which can be boring - about fifty each is usually enough. It also means that, if you have one poor interviewer, this affects the results badly, since each interviewer does a high proportion of the interviews. The disadvantage of having a large number of interviewers is that it requires more effort to recruit, train and supervise them.

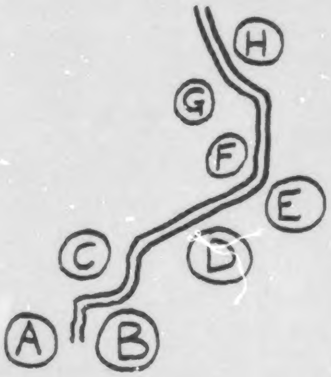
Apart from the practical advantage of economising on transport costs by having the interviewers travel as a team, there is a technical advantage also; it enables you to see if there were differences between the results obtained by different interviewers. If the interviewers do roughly similar shares of the interviewing in each place they visit, their results should be similar, and you can make use of this fact to check on their work. I'll explain how in Chapter 8.

Training interviewers

The interviewer's job falls roughly into two parts - first, locating the respondent and gaining his cooperation, then conducting the interview. The researcher has to work out in detail exactly how he wants the interviewers to carry out these tasks. It is a good idea to write down these instructions and give a copy to each interviewer. Bring the interviewers together for a training course and go through the instructions thoroughly.

If you want the interviewers to do any sampling on the spot, you must tell them exactly how to proceed. Some of the instructions for one of LDTC's surveys were as follows:

> Take the first house that you come to as you apprach the village. Toss a coin. If it is heads, take that as your first house. If it is tails, take the next house as your first house.
> When you have visited your first house, decide which two houses are nearest to that house. Choose the further of the two. When you have visited that house, decide which two houses are nearest to it, not counting any of the houses you have considered already. Choose the further of the two. Carry on like this.



> Example. (A) is the first house you come to. You toss a coin. It's tails, so you take (B) as your first house to visit. (C) and (D) are the nearest houses to (B), so you take the further one - (D). The nearest to (D), not counting (A), (B) or (C) are (E) and (F). You take the the further one - (F). The nearest to (F), not counting (E), are (G) and (H). You take (H).

If the interviewers then have to decide who to interview within the household, they need clear instructions for this also, such as, 'Interview everyone aged between 13 and 20 inclusive,' or 'Interview any mother who has a child aged five or under.'

Decide what you want them to say when they first come to a household, and get them to practise it. It could be something like this:

> 'Good morning. My name is .... I work for an organisation called .... We are conducting a survey about ....
> I wonder if I could speak to the head of the household.'

The interviewer can go into detail, if necessary, about the organisation, the reason for the survey, the authorisation for it, the confidentiality of the information and so on.

One of the problems with surveys is that some people are never at home when the interviewer calls - the problem of 'non-contact' - and another is that some people refuse to be interviewed - the problem of 'non-response'. In order for you to assess, at a later stage, how reliable the results of the survey are, it is essential that the interviewers keep a record of these difficulties. One way to do this is to prepare special forms, sometimes called 'Failure forms'. For every household where the interviewers fail to get an interview, for whatever reason, they fill in one of these forms and bring them back with the completed questionnaires. If possible, they should record some details about the people they have failed to interview, such as their sex and an estimate of their age.

Again, you must give the interviewers precise instructions about what to do. Here is another quotation from the interviewers' instructions for a survey by LDTC:

> If you visit a house when there is no one present, or no senior member to talk to, make a note of this house to remind yourself to call back. You must visit a house at least three times before you give up.
>
> It may happen that you cannot enter a household - because of a recent birth, for example - or that the head of the household is never at home when you call, or that he/she keeps asking you to come another time or that he/she refuses to cooperate. In other words, you may be unable to get an interview. If this happens, for any reason at all, fill in one of the Failure forms (the ones on pink paper). Do not interview someone else instead.

If the interviewers are to visit villages, they need to know quite a lot about the distance-teaching institution in general and about this survey in particular, so you should include a talk on this in the training course. In Lesotho, the interviewers have to visit the village chief first and he might ask many questions. (A chief acts rather like a head of household for the whole village; it would be quite wrong to start visiting people without his consent.) After that, the chief might call a meeting of the villagers and they too might ask the interviewers many questions. If the interviewers travel as a team, you should appoint one interviewer to act as the team leader in talking to the chief and answering questions.

When it comes to conducting the interview, I favour giving the interviewer as much help as possible on the questionnaire itself by writing out the preamble at the beginning, writing the questions exactly as you want them to be read out, giving clear branching instructions, and reminding the interviewer to say 'Thank you' at the end. The interviewers then just need to be trained to follow the questionnaires accurately. You should also stress that they must not influence people's answers by saying 'Right' or 'Wrong' or 'Good' or by expressing surprise or horror at what people say; they should just ask the questions in a non-commital tone of voice and record the answers.

A good way to train interviewers how to conduct an interview is to fill in a dummy questionnaire showing the answers that a typical respondent might give. The instructor then pretends to be this typical respondent; the trainee interviewer interviews the instructor who gives the answers that are marked on the dummy questionnaire, trying to make the interview as lifelike as possible. The trainee does the full interview, filling in a questionnaire. At the end, the instructor comments on the trainee's technique and also compares the two questionnaires to see if the trainee has recorded the answers correctly. You can prepare several different dummy questionnaires, of course, so as to go through this procedure several times with each trainee.

Another simple training device is to encourage the trainees to do a few practice interviews with their family or friends, so as to become more familiar with the questionnaire and to gain confidence.

In the preamble to the questionnaire the interviewers promise to treat the respondents' answers in confidence, and it is important that they should keep this promise. To impress the importance of this on the interviewers, you may insist that, as a condition of their employment, they sign a declaration promising not to divulge information given in the interviews to anyone outside the survey team.

Clearly, there is a lot to do in a training course for interviewers, so you should allow plenty of time for it, perhaps as much as two or three days. It is sometimes tempting to do the training hastily in the hope that the interviewers will pick up the job as they go along, but this is a mistake. An interviewer who is not fully trained is likely to make the same errors right through the fieldwork.

The pilot survey

If you are going to spend a lot of time and effort on a survey, it may be worthwhile to do a pilot. This is a small scale version of the full survey. For example, if you were planning to interview 240 people in eight villages, the pilot might consist of 30 interviews in one village. The pilot is a pre-test of the survey procedures and of the questionnaire. The purpose is to find out if there is anything wrong with the instructions to the interviewers or with the questionnaire before you proceed with the full survey.

In a pilot survey, the role of the interviewers is a bit different from usual. As well as recording people's answers, they should also note if the respondents seem to misunderstand any of the questions. And they do not have to stick so closely to the exact wording of questions on the questionnaire. If they think that a different wording would be clearer, they should try it out and report back to the researcher. This requires more skill than interviewing in the main survey so, if possible, you should use experienced interviewers on a pilot survey.

It is worthwhile to count up and tabulate the results of the pilot, even though you have interviewed only a small number of people. This will show up any errors in the branching instructions (see the section on branching in Chapter 5). If there are any strange results, these might indicate faults in the questionnaire, perhaps in the translation.

Researchers sometimes add the results of the pilot to the results of the main survey, to increase the sample size. There is no firm rule about this. It depends partly on how many changes you make to the questionnaire in the light of the pilot, and partly on the sampling method you use. If the main survey questionnaire is not very different from the pilot and if the inclusion of the pilot respondents will not distort the sample, you may add in the pilot results, but otherwise you shouldn't.

Administration

Before recruiting the interviewers, you should write out their conditions of service. You must decide, for example, whether to pay them extra for working outside office hours or at weekends (both of which might be necessary), and how much to pay for nights spent away from home. If the interviewers are being employed solely for the period of the fieldwork, make this quite clear at the outset.

If the survey involves visiting remote villages, you may have to make special arrangements for travelling and camping. If you allow the interviewers to arrange their own transport and accommodation on expenses, you must decide in advance how much they are allowed to spend. Check that the interviewers are insured. If they are temporary employees, you might have to alter the institution's regular insurance policy to cover them.

With rural surveys in Lesotho, it is necessary to send letters in advance to the chiefs of the sample villages, signed by a senior civil servant. The interviewers' leader should take copies of these letters. We have also found it useful for the interviewers to take small cards printed with the name and address of the LDTC. When an interviewer visits a household, he signs his name on the card and gives it to the head of the household.

During the interviewing, the leader should sit in on a few of his colleagues' interviews, if possible, to check that all the interviewers are handling it in the same way. It is best to let them do a few on their own first, to gain confidence, before making this check. He should also check a few of their completed questionnaires to see if they are recording answers and following branching instructions correctly.

Do not wait till the end of the fieldwork before you take in the questionnaires. Take in each batch of questionnaires and failure forms as they are completed. Check that they are being filled in correctly. If it is a large survey, you can start processing the results of the first batch while the fieldwork is still going on. Discuss with the interviewers any problems they seem to be having. See if any interviewer is getting a particularly high number of refusals and, if so, see whether his approach to people can be improved. (If an interviewer is doing very badly, you might take him off the job altogether.) At some point you should put a serial number of each questionnaire, i.e. number 001 for the first questionnaire, number 002 for the second, and so on. (These serial numbers are useful later in the analysis.) You can arrange for the interviewers to do this during the fieldwork or you can do it as part of checking in the questionnaires. Needless to say, you should take great care not to lose any questionnaires.

Some organisations pay interviewers at so-much-per-interview, others at so-much-per-hour. I prefer the second system. The first one encourages interviewers to depart from the strict sampling rules; if they are having difficulty in finding anyone at home at the selected house, they will be tempted to take someone else instead. The second system, by contrast, pays them equally for all the work they do - revisiting households as well as doing interviews. If a team of interviewers go off together to a village and return together some days later, the simplest payment system is so-much-per-day.

Payment systems, camping equipment, expenses claims and the like might seem to be tedious matters for a researcher to have to attend to, but it is important that he should. The researcher's attitude to the survey communicates itself to the interviewers. If he appears not to care very much about how things get done, the interviewers won't care either.

78

7 Processing the data

This chapter describes several ways of taking the data from the questionnaires and turning it into tables of figures.

Counting straight from questionnaires With small surveys and simple analysis you can count the results straight from the questionnaires.

Coding onto transfer cards Answers are transferred from the questionnaires onto transfer cards, usually in the form of code numbers. These cards make it easier to count up the answers.

Edge-clipped cards Using a line of holes along the edge of the transfer cards speeds up the sorting of the cards.

Squared paper Putting all the answers, in coded form, onto a large sheet of squared paper is another simple method for processing data from a small survey.

Using a computer Putting data into a computer using 80-column cards. Programming with SPSS. Using a counter-sorter.

Which method to use? Use the least sophisticated method that is feasible for the job you want to do.

The survey of poultry-keeping, which I used as an example in earlier chapters, set out to answer such questions as, 'How many households keep poultry?' 'How many birds do they keep?' 'What sort of food do they give to the birds?' and so on. The information with which to answer these questions is contained in the pile of completed questionnaires, but not in a form that is easy to get at. Eventually, the results of the survey will be presented in statements like these: 'Fifty-two per cent of rural households have poultry.' 'Four-fifths of poultry owners have fewer than five birds.' 'It is women rather than men who take care of the birds.' 'Those with large holdings of poultry (over 200 birds) tend to be people who have had a higher level of education.' And so on. Rather as a factory takes in raw materials and turns them into finished goods, a research team takes in the 'raw data' - the information contained in the questionnaires - and converts it into tables of figures. There are several ways of doing this. The method you choose will depend on the amount of data you have and the complexity of the analysis that you want to perform.*

---

* There is some uncertainty over the use of the word 'data'. The word comes from Latin where it means, simply, 'things that have been given'. Some people use it as a plural in English, as it is in Latin, on the grounds that it refers to all the separate items of information that respondents have given; they would say, for example, 'Many data were collected but only a few of them were used in the analysis.' Other people, including me, feel that it is more natural in English to use 'data' as a singular, like the word 'information'; they would say, 'Much data was collected but only a little of it was used in the analysis.' And other people use it sometimes one way and sometimes the other. Those who use it as a plural tend to think that they are right and others are wrong. I think it is a matter of taste.

Counting straight from questionnaires

If the questionnaires are short and there are not too many of them (no more than 60, say), you can count the answers straight off the questionnaires. For example, if the answers to question 1 were Yes, No or Don't know, you could sort the questionnaires into three piles according to their answers to this question - one pile for those who answered Yes, one for those who answered No, and one for those who said they didn't know. Then you would count the questionnaires in each pile.

Alternatively you could count the answers by the 'tally-mark' method. You begin by drawing up a little table on a sheet of paper, like this:

Question 1	Yes	
	No	
	Don't know	

Then you put a mark for each questionnaire in the appropriate part of the table. Say you had 38 questionnaires in all; after you had tally-marked all of them, the table might look like this:

Question 1	Yes	|||||||||||||
	No	||||||||||||||||||||||
	Don't know	|||

A slight improvement on this is to mark every fifth one by drawing a line across the previous four marks, thus grouping them in fives. If you had used this method, the same table would look like this:

Question 1	Yes	𝍸 𝍸 |||
	No	𝍸 𝍸 𝍸 𝍸 ||
	Don't know	|||

You then count up the tally-marks, to find that 13 respondents said Yes, 22 said No, and three said Don't know, making a total of 38.

With some questions, you will find an extra category is necessary - 'Inadequate information'. This is for when the interviewer missed out a question by mistake, or forgot to record the answer, or made some ambiguous mark on the questionnaire which no-one can read. It is better to have a category for these, rather than to omit them, so that you always make one tally-mark for each questionnaire. You then know that the tally-marks should add up to the total number of questionnaires you have. If they do not add up correctly, you know you have made a mistake. Where respondents have skipped questions, because of the branching instructions, you will need another category - 'Does not apply'. The tally-mark table for another question might look like this:

Question 14	Yes	𝍸 𝍸 𝍸 𝍷𝍷𝍷
	No	𝍸
	Don't know	𝍷
	Does not apply	𝍸 𝍸 𝍷𝍷
	Inadequate info.	𝍷𝍷

When you go through all the questionnaires like this, counting the answers to one question, this is called a 'frequency count'. With small surveys, the frequency counts may be all you want. A set of frequency counts from a survey would look like this:

Question 1. Did you hear any radio announcements last week about road safety?	Yes	15
	No	18
	Not sure	5
Question 2. Did you hear only one or more than one?	Only one	2
	More than one	11
	Not sure	1
	Does not apply	23
	Inadequate info.	1

And so on, for each question. It is often convenient to record these figures on an unused questionnaire.

If frequency counts are all you want, the quickest way to do the counting is to go through the questionnaires using the tally-mark method I have just described. However, if you are also going to want cross-tabulations, this method is very unwieldy.

A cross-tabulation (often abbreviated to 'crosstab') is a table which divides up the respondents and shows the answers of each group separately. For example, you might want to compare the answers of men and women to a questionnaire on family planning. Say the sample contained 84 men and 92 women; the crosstab for one question might look like this:

Question 29. Have you heard of the contraceptive pill?		Men	Women
	Yes	12	29
	No	44	35
	Not sure	23	25
	Inadequate info.	5	3
	Total	84	92

This table means that, of the 84 men who were interviewed, 12 said Yes to question 29, 44 said No, 23 were not sure and 5 did not have their answers recorded properly; of the 92 women who were interviewed, 29 said Yes, 35 said No, 25 were not sure and 3 did not have their answers recorded properly.

To produce this table straight from the questionnaires would be tedious. You would take each questionnaire in turn; you would find the page where the respondent's sex was recorded; then you would turn to question 29 to see what answer the respondent gave and you would make a mark in the appropriate column (i.e. under 'Men' or 'Women'). And you would have to do this 176 times, just to produce this table. In analysing a survey, you might want dozens of cross-tabulations.

Coding onto transfer cards

In such cases it is more convenient to transfer the information from each questionnaire onto a single sheet of paper or a card called a 'transfer card'. (I prefer cards to sheets of paper since they will be handled a lot, and sheets of paper easily get creased or torn.) The transfer cards for the poultry survey might look like this:

Lesotho Distance Teaching Centre

POULTRY SURVEY Transfer Card

Serial number ☐☐☐

Question 1		11		21		31		41 (sex)	
2		12		22		32			
3		13		23		33		42 (age)	
4		14		24		34			
5		15		25		35			
6		16		26		36		43 (educ)	
7		17		27		37			
8		18		28		38		44 (village)	
9		19		29		39			
10		20		30		40			

You design the transfer card so that it will contain all the information from the questionnaire, and you have a number of them printed - one for each completed questionnaire.

You might do the transferring yourself, but let's say you give this job to an assistant. He takes questionnaire number 001; he takes a transfer card and writes 001 in the boxes marked 'serial number'. Then he looks at the answer to question 1. Let's say there are only three possible answers to question 1 - Yes, No and Don't know - and that this respondent has answered Yes. Your assistant could write the word 'Yes' in the box for question 1 on the transfer card, but it is easier to convert the answers into code-numbers and to write numbers in the boxes. You could decide to have the answer 'Yes' represented by code-number 1, 'No' by code number 2, and 'Don't know' by code-number 9. The respondent said Yes to question 1, so your assistant writes '1' in the box for question 1 on this respondent's transfer card. This procedure of converting answers into code-numbers is called 'coding' and the people who do it are called 'coders'.

In order to do their job, the coders obviously need to know what code numbers you want them to use. The coding instructions for question 1 in this example are very simple:

Question 1.	Yes	Code 1
	No	Code 2
	Don't know	Code 9

If all the questions were of the Yes/No type, you could use these instructions for every question, with minor additions for 'Inadequate information' and 'Does not apply':

For all questions.	Yes	Code 1
	No	Code 2
	Don't know	Code 9
	Inadequate information	Code 0
	Does not apply	Leave the box blank

If any questions had different answers from these, you would need special instructions for those questions, such as the following:

Question 2. ('How many birds do you have?')

One or two birds	Code 1
Three or four birds	Code 2
Five to ten birds	Code 3
Eleven to twenty birds	Code 4
Over twenty birds	Code 5
Don't know	Code 9
Inadequate information	Code 0
Does not apply	Leave box blank

The complete set of coding instructions for this questionnaire, which is known as a 'coding frame', would begin like this:

Lesotho Distance Teaching Centre. POULTRY SURVEY.
Coding Frame.

Throughout the questionnaire, use 9 for Don't know
0 for Inadequate information
and leave box blank for Does not apply.

For Yes/No answers, use 1 for Yes
2 for No

Other questions as follows:

Question 2:	1 or 2 birds	Code 1
	3 or 4 birds	Code 2
	5-10 birds	Code 3
	11-20 birds	Code 4
	Over 20 birds	Code 5
Question 5:	Feeds them	Code 1
	Fend for themselves	Code 2

Returning now to respondent number 001, suppose the first five answers on his questionnaire looked like this:

1. Do you have any poultry?

   Yes ☑ No ☐ → Q24

2. How many birds do you have?

   1 or 2 ☐
   3 or 4 ☑
   5 to 10 ☐
   11 to 20 ☐
   Over 20 ☐ → Q19
   Not sure ☐

3. Have you ever had them vaccinated?

   Yes ☐
   No ☑ → Q5
   Can't remember ☐ → Q5

4. Were they vaccinated by an agricultural demonstrator?

   Yes ☐ No ☐ Not sure ☐

5. Do you feed them or do they fend for themselves?

   Feeds them ☐
   Fend for themselves ☑
   Other ☐ (Explain ....)

After the coder had transferred just these five answers to the transfer card, it would look like this:

Lesotho Distance Teaching Centre
POULTRY SURVEY Transfer Card

Serial number 001

Question 1	1	11		21		31		41 (sex)	
2	2	12		22		32			
3	2	13		23		33		42 (age)	
4		14		24		34			
5	2	15		25		35		43 (educ)	
6		16		26		36			
7		17		27		37		44 (village)	
8		18		28		38			
9		19		29		39			
10		20		30		40			

By means of coding, a person's answers to a long questionnaire can be condensed onto a single transfer card. Having made these cards - one for each respondent - you then do all the counting and cross-tabulating from the cards. This is much quicker than handling the actual questionnaires.

If you look back at the section on 'Pre-coding' in Chapter 5 (page 56) you will appreciate that pre-coding the answers when writing the questionnaire makes it very easy later to transfer the answers onto transfer cards. For those questions that are pre-coded, you do not have to bother with a coding frame.

You might remember that I included a space for the respondent's serial number on page 1 of the questionnaire (Chapter 5, page 49). This is needed if you transfer the information onto a card. Suppose that you notice, at some point in the counting, that one card seems a bit strange; for example, the answers, as coded on that card, indicate that the respondent is a poor woman yet that she has a commercial farm of 500 birds. If you want to check the card against the questionnaire to make sure the card was filled in correctly, you have to know which questionnaire the card refers to. This is where the respondent's serial number comes in. You give a number to each questionnaire (e.g. starting with 001 and going on from there) and you also put this number onto the transfer card. If the strange card is number 147, you just pull out questionnaire number 147 to check the card against it.

You should print the title of the survey on the transfer cards. If you are working on several surveys at the same time, it might be helpful to use different coloured card for each survey. Needless to say, you should keep all the transfer cards for one survey together;

perhaps bind them with an elastic band or keep them in a box when you are not using them.

Open questions (see 'Closed and open questions', Chapter 5) create work at the coding stage. Suppose you had included a question for infertile women in a family planning survey - 'Who have you asked for advice on this problem?' First, you would go through about 30 of their questionnaires, jotting down their answers to this question. Then you would devise a set of categories to put these answers into, and give a code number to each. It might look like this:

Hospital doctor	Code 1
Clinic nurse	Code 2
Traditional healer ('witch doctor')	Code 3
Mother-in-law	Code 4
Other female relative	Code 5
Clergyman, priest	Code 6
Other	Code 7

Then the coders would put each woman's answer to this question into one of these categories, marking the appropriate code number on the transfer card. This is called 'post-coding'. ('Pre-coding' just means 'coding before the fieldwork', and 'post-coding' means 'coding after the fieldwork'.)

A minor problem in coding is when people give more than one answer to a question. In the example I have just given, a woman might have consulted a doctor and a priest. In this case you would code her as '1,6'. This is called 'multi-coding'. Obviously if each person can give more than one answer, the number of answers to that question will not add up to exactly the number of respondents.

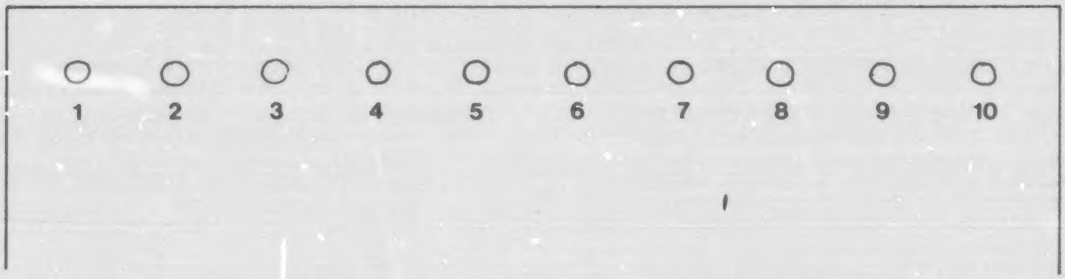
It is important that the information is transferred correctly from the questionnaire to the transfer cards. To guard against error, you might have the questionnaire transferred onto two sets of cards by two different people. You would then compare their cards and investigate any discrepancies between the two sets. Discrepancies are most likely to occur with post-coded answers.

Edge-clipped cards

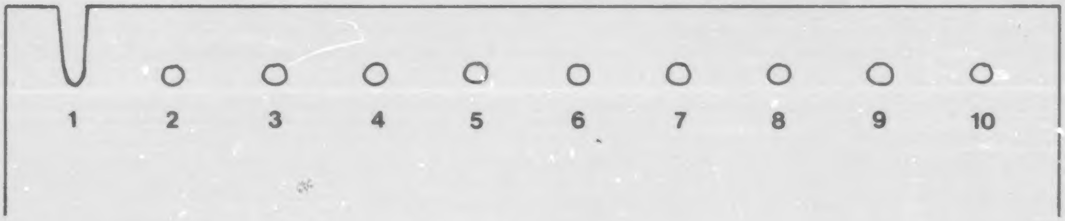
In the last section, I did not explain in detail how you would use the transfer cards to produce a crosstab. Take the family planning example again. You have 176 transfer cards, and you want to compare the answers of men and women to question 29. Let's say that the sex of the respondent is recorded on the transfer card as question 57, with code 1 for men and code 2 for women.

The quickest way to use the cards is first to go through them all just looking at question 57 and sorting the cards into two piles - men and women. Then you would go through the men's pile and tally-mark their answers to question 29. Finally you would go through the women's pile, tally-marking question 29. This method breaks the whole operation into two stages; first, you sort the cards into piles, then you count the answers of each pile.

Edge-clipping is a way of making the first stage of this operation (the sorting) much quicker. In order to use edge clipping, you need to have a line of holes along the edge of the transfer card, like this:

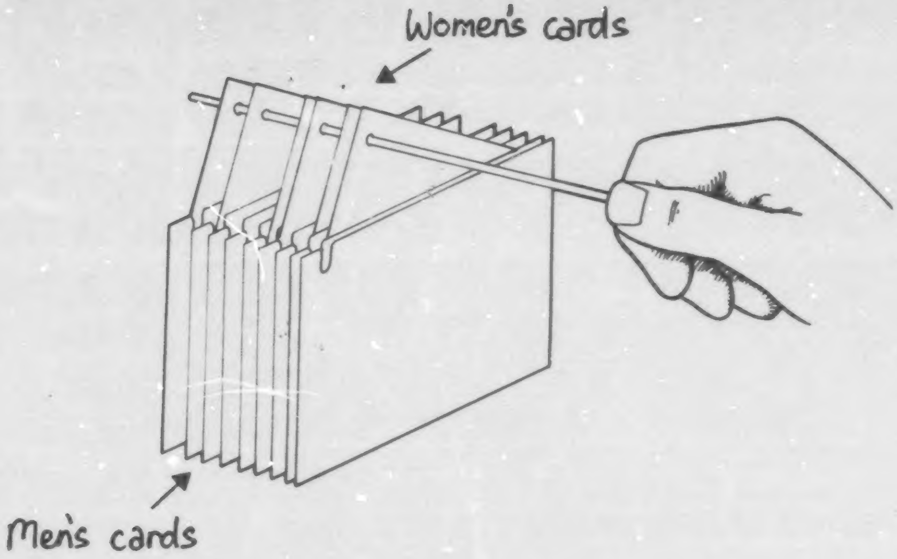


Let's say you decide to use hole number 1 to record the respondent's sex. For men, you could clip away that part of the card between the hole and the edge, like this:



You can obtain a special tool for this, or you can use scissors. For women, you would leave the card alone.

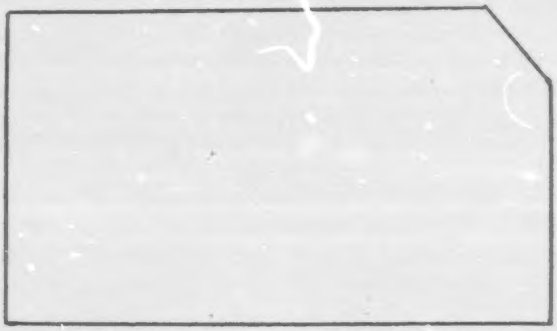
Now, imagine you have all your 176 cards together, with the men's cards and the women's cards jumbled up. In order to separate the men from the women, you poke a knitting needle or a skewer through hole number 1, so that it goes through hole number 1 on every card. If you now raise the knitting needle, all the women's cards will remain on the needle, while all the men's cards will fall onto the table. I have shown only a few of the 176 cards in this diagram, to make it clearer.



In this way, you can sort the cards into two piles - men and women - very quickly.

If you are going to use the cards a lot, it is a good idea to strengthen the edge with tape before you do the clipping (but be careful not to cover any of the holes).

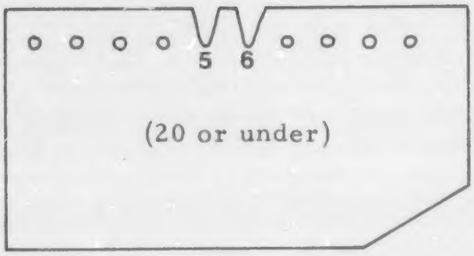
Obviously, you must have all the cards with their holes at the top and their front sides facing you before you push the needle through. (If one card was facing the other way, the needle would go through hole number 10 on that card instead of hole number 1.) You can make sure of this by taking off one corner of the transfer card, so that all the cards are this shape:

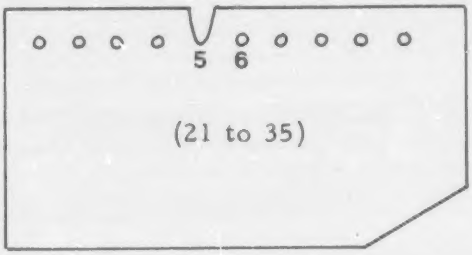


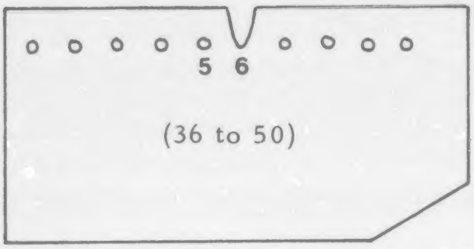
With cards this shape, you can see at a glance if any card in the pack is facing the wrong way.

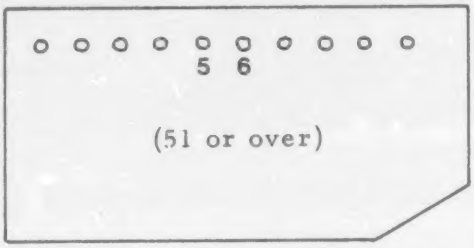
You can use edge-clipping even when the respondents fall into more than two categories. Suppose you have divided up the respondents into these age-groups: 20 or under, 21 to 35, 36 to 50, 51 or over. You can use edge-clipping for this, but you will need two holes.

Let's say you use hole number 5 and hole number 6. With hole number 5, you could separate those 35 or under from those 36 or over; let's say you clip hole 5 for the younger respondents (35 or under) and leave it alone for the older ones. Now you take the younger group only. With hole number 6, you could separate the people 20 or under (clip hole 6) from those 21 to 35 (leave hole 6). Now take the older group (36 or over). For them, hole 6 could separate the people 36 to 50 (clip hole 6) from those 51 or over (leave hole 6). The four groups are now represented on holes 5 and 6 as follows:

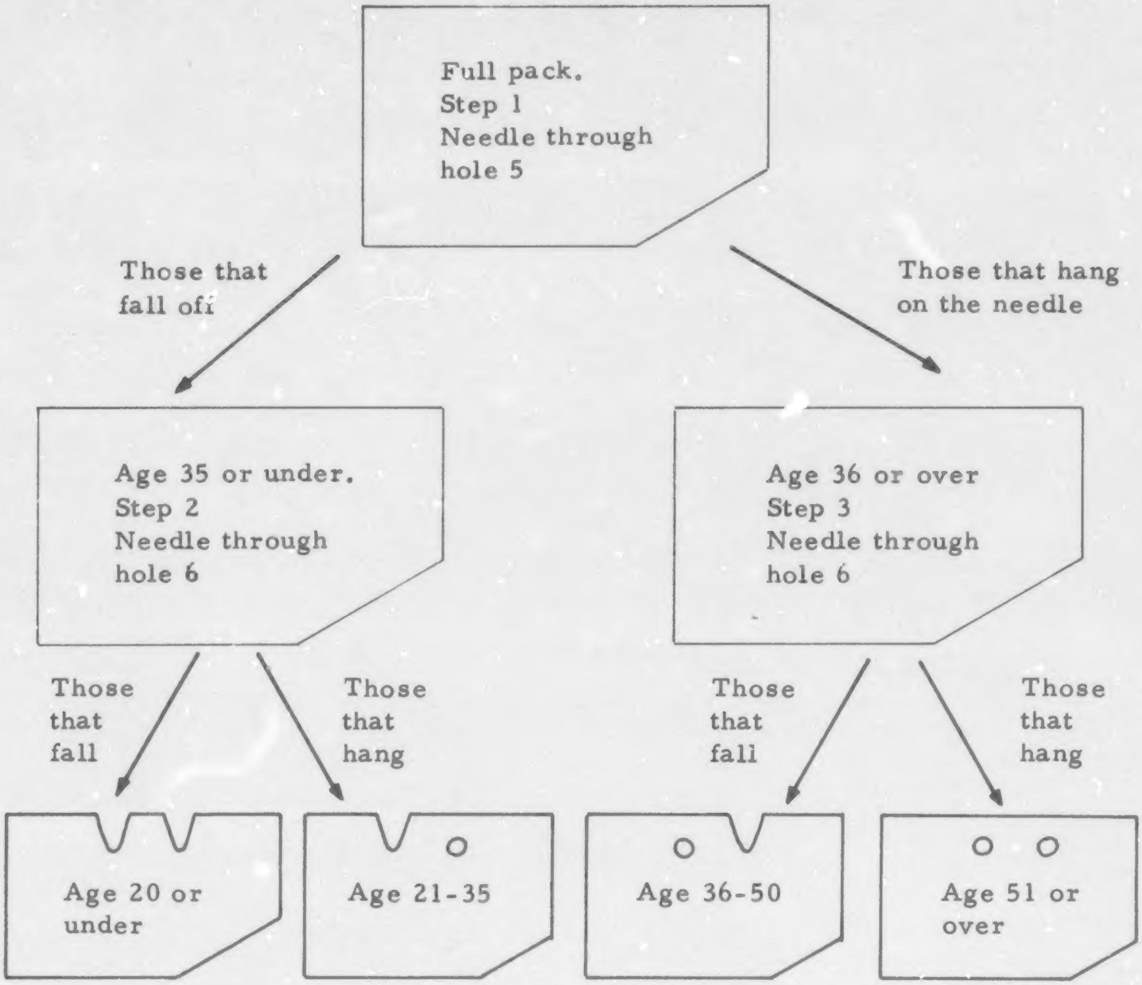








To separate the full pack into these four piles, you need to use the needle three times, as in this diagram:


Full pack.
Step 1
Needle through
hole 5
Those that
fall off
Those that hang
on the needle
Age 35 or under.
Step 2
Needle through
hole 6
Age 36 or over
Step 3
Needle through
hole 6
Those
that
fall
Those
that
hang
Those
that
fall
Those
that
hang
Age 20 or
under
Age 21-35
Age 36-50
Age 51 or
over

In principle you could use edge-clipping for any number of categories. (With three holes, you can divide the pack into eight piles and so on.) But in practice it becomes difficult to keep track of what you are doing if you have a complicated system of edge-clipping. It is best to use edge-clipping just for a few basic things like respondent's sex, age and education.

Squared paper

Large sheets of squared paper are another way of handling the data. Say you had 60 questionnaires, with 35 questions in each; you would prepare a large sheet of paper (say four times the size of the standard A4 sheet) divided into 35 columns and 60 rows, like this:

	Question								
	1	2	3	4	5	6	7	and so on up to	35
respondent 01								-- --	
02								-- --	
03								-- --	
04								-- --	
and so on down to								-- --	
60								-- --	

You convert the answers into code numbers, in the same way as for transfer cards, but you write the code numbers on this large sheet of paper. First you take the questionnaire of respondent 01; you code his answer to question 1 and write the number in the top left-hand square; then you code his answer to question 2 in the next square along, and so on until you have transferred all the answers from his questionnaire to the top line of the sheet. Then you do the same for respondent 02 on the second line.

Like transfer cards, this method makes it easier to produce crosstabs. Let's say that the survey was about schoolteachers' use of educational aids; question 4 records whether or not the teacher had listened to a schools radio programme in the week before the interview, and question 31 records whether or not the teacher had a formal teaching qualification. The coding for question 4 is: Yes - 1, No - 2, inadequate information - 0; for question 31, it is: qualified - 1, unqualified - 2, inadequate information - 0. You want to know whether qualified teachers make more use, or less use, of schools' broadcasts than unqualified ones. So you need a crosstab, i.e. you want to divide teachers into qualified and unqualified, and then count up separately their answers to question 4. You could describe this crosstab as 'Question 4 by question 31' or as 'Whether listened to schools' broadcast by qualification'.

First you make yourself a table, like this:

		Question 31		
		Code 1	2	0
Question 4	Code 1			
	2			
	0			

Then you take your sheet of squared paper, onto which you have already transferred all the answers from the questionnaire, in code form. You are interested only in questions 4 and 31. It is useful to put rulers alongside the columns you are interested in, to prevent yourself from reading the wrong column. Taking the top line first (i.e. respondent 01) you might find he is recorded as '1' for question 31 and also '1' for question 4. So you put a tally-mark in the top left-hand box of your table. The following diagram shows the first seven lines of the squared sheet (ignoring all columns except 4 and 31) and also how the table would look after you had tally-marked just these first seven:

	Question	
	4	31
Respondent 01	1	1
02	2	1
03	2	2
04	2	2
05	2	0
06	1	1
07	1	1

		Question 31		
		Code 1	2	0
Question 4	Code 1	III		
	2	I	II	I
	0			

After you have tally-marked all sixty, you count up the tally-marks, which give you the crosstab you wanted. Perhaps it would look like this:

		Question 31. Teaching qualification		
		Qualified	Not qualified	Inad. info.
Question 4 Schools broadcast previous week	Yes	21	7	0
	No	16	10	1
	Inadequate info.	1	2	2

When you produce a crosstab by any tally-mark method, you should add up all the numbers to make sure you have the right total. In this example there ought to be 60 respondents in all.

For small samples and short questionnaires, the squared paper method is quite good. But it has its limitations. If you have more than about 100 questionnaires and more than about 60 questions, the sheet of squared paper becomes unmanageably large. If you make the rows and columns very narrow, you are more likely to make mistakes in transferring information onto the sheet or in reading the information from it. Also, it is difficult to do more complicated analysis by this method.

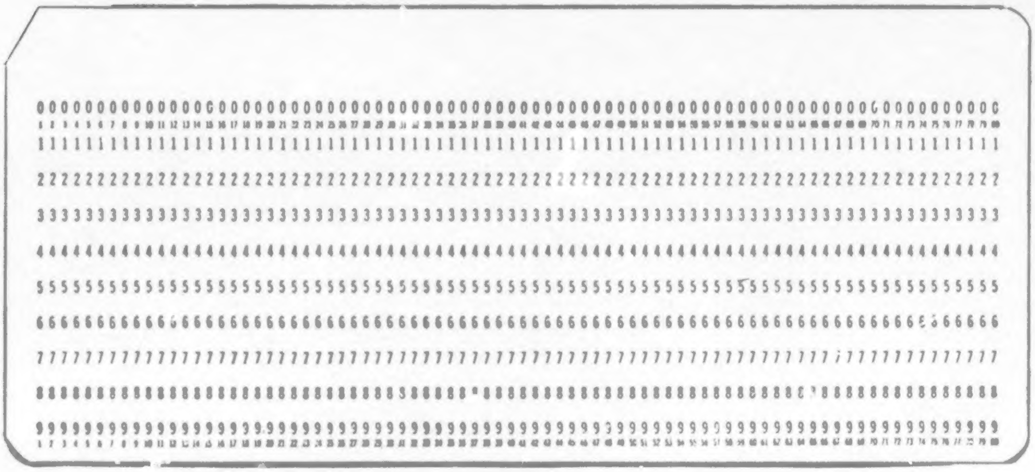
Using a computer

I won't attempt to tell you all you need to know in order to process survey data on a computer. The topic is too big and complex for me to tackle in this book, and computer technology is changing so fast that if I wrote about it at length, much of the advice I gave would be out-of-date in a few years. But computers are becoming smaller, cheaper and more widely available, and they can process data very efficiently. For large amounts of data or for complicated analysis, they are essential. So I will say a little about computers, for the benefit of those readers who have access to a computer and who are thinking of using it. If you do not have access to a computer, skip this section.

All the methods of data-processing that I have described so far in this chapter require a person to do repetitive clerical work. Suppose you are analysing the results of the poultry survey and you are using transfer cards. You want to see whether there is a difference between mountain people and lowland people in the way they feed their birds. You give the pile of transfer cards to your assistant and tell him you want a crosstab of item 5 on the transfer card (i.e. answers to the question 'Do you feed them or do they fend for themselves?') by item 44 ('Village'). He goes through the cards one by one, using the tally-mark method, and eventually brings back the crosstab you wanted.

When you use a computer to process survey data, the computer takes the place of the assistant, and you go through the same steps with a computer that you do with an assistant. Just as you give the survey data to your assistant, in the form of the transfer cards, so you have to feed the data into the computer. Then, like telling your assistant what you want him to do, you have to give instructions to the computer.

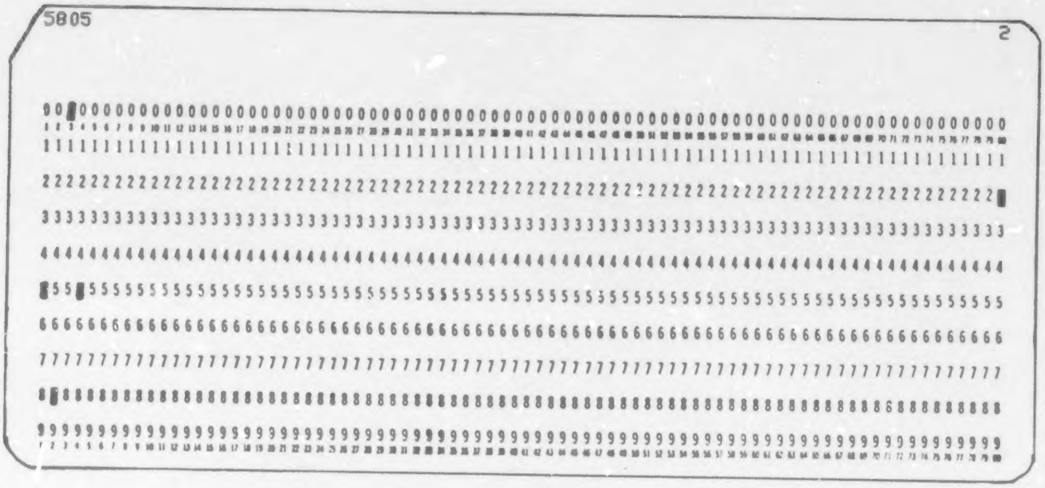
The data and instructions are what you put into the computer, so they are known as 'input'. Computers can receive input in several different ways; the system you use will depend on the facilities available with the computer you are using. One method of putting data into the computer - probably the most widely used method at the moment - is to put the data on specially designed cards and to feed these cards into a device called a 'card-reader'; the card-reader reads the data from the cards and passes it on to the computer in a form that the computer can accept. The transfer cards I described earlier were designed to be read, sorted and counted by people; computer cards are just a special kind of transfer card which are intended to be read by a card-reader. One of these cards looks like this:



You put information on the card by making small rectangular holes in it. The holes are actually made by a special machine called a 'key-punch', and the job is usually done by a person, called a 'card-puncher', who is skilled at using this machine. If you gave the card-puncher these instructions:

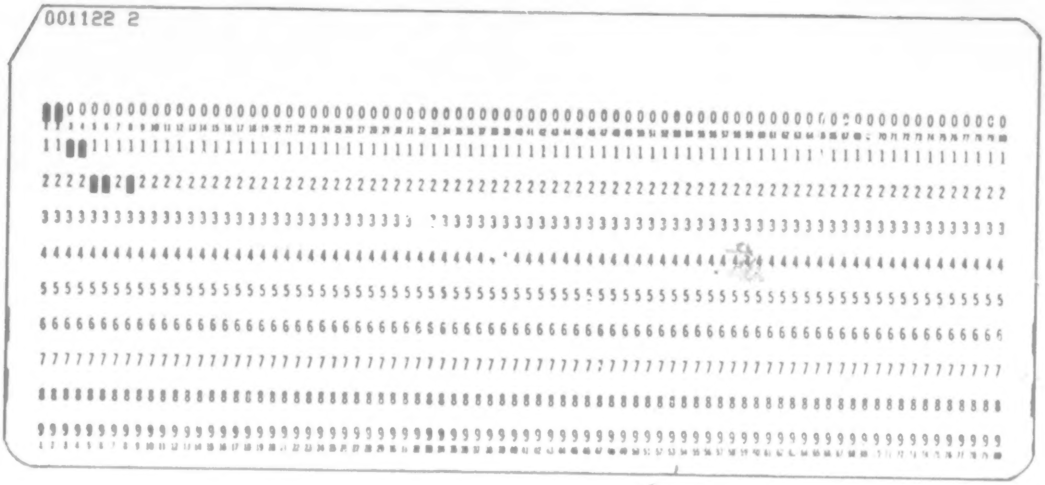
Column 1	Code 5
" 2	" 8
" 3	" 0
" 4	" 5
" 5 - 79	Skip
" 80	Code 2

you would get a card looking like this:



The machine on which this particular card was punched types the numbers along the top as well as punching the holes. This makes it easier for people to see what's punched on the card, but it makes no difference to the computer; the card-reader only looks at the holes.

To put the information from a questionnaire onto a computer card, you have to convert all the answers into code numbers and then put the answer to each question onto a column of the computer card. Take the poultry example again. If each respondent had a three-digit serial number, you could put this number on columns 1 to 3 of the card. Then you could put his answer to question 1 on column 4, his answer to question 2 on column 5 and so on. If the respondent whose questionnaire is given on page 80 had serial number 001, the first eight columns of his computer card would look like this:

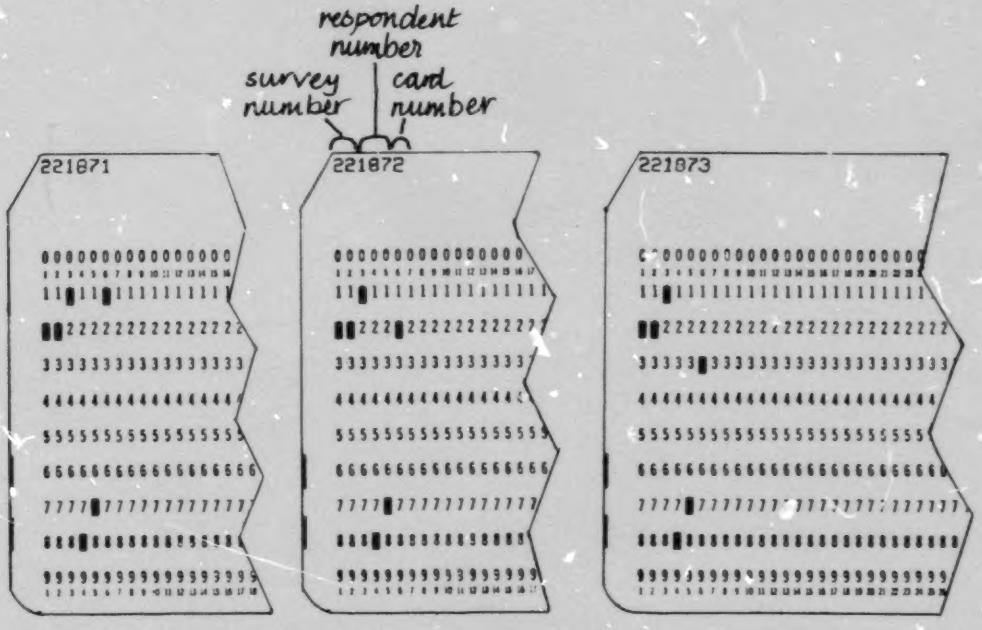


This shows his serial number -001 - on columns 1 to 3, and then his answers to the first five questions, converted into number form (as on page 81), on columns 4 to 8.

If you use a computer for several different surveys, you should also put on the card an identification number for the survey, e.g. 01 for the poultry survey, 02 for the family planning and so on. This is because computer cards all look alike. It would be disastrous to get the cards from one survey mixed up with those from another. If each survey had its own code number, you could at least sort them out again.

A computer card has 80 columns. If you have more than 80 items of information - including survey number, respondent's serial number, and then all the questions - you might need to go on to a second card. (Some small computers cannot take this, so be sure to find out in advance.) If you do have two or more cards per respondent, remember to put the survey number and the respondent's serial number on every card, and also to set aside a column for the card number. (Though you can have more than one card per questionnaire, you never put more than one questionnaire on one card.)

As an illustration, the following shows the first six columns of the cards of respondent 187, in survey 22, in which each respondent has three cards:



If you are going to use a computer to process the survey results, it is best to take this decision very early and to have a talk with the person in charge of the computer. It may be possible to design the questionnaire so that the card-puncher can read the information in number form straight from the questionnaire and punch it onto cards. This requires that you specify the card and column numbers on the questionnaire; the top right corner of the first page might look like this:

Card 1, Col	1	2	3	4	5	6
	2	2				1

Before the questionnaires went to the card-puncher, you would fill in the respondent's serial number. Given these instructions, the card-puncher would begin the first card with the correct six digits (survey number in columns 1 and 2, respondent's serial number in columns 3 to 5, card number in column 6.)

The rest of the questionnaire would also have to be laid out to make things clear for the card-puncher, perhaps like this:

		(Card 1, Col. 7)	
1. Do you have any poultry?	Yes	1	
	No	2 --------	→ Q24
2. How many birds do you have?		(Col. 8)	
	1 or 2	1	
	3 or 4	2	
	5 to 10	3	
	11 to 20	4	
	over 20	5 --------	→ Q19
	not sure	9	

The numbers in brackets are the card and column numbers, for the card-puncher. The numbers not in brackets are the code numbers that the interviewer rings and the card-puncher punches. The question numbers to the right are the interviewer's branching instructions. At the beginning of the second card, the following instructions to the card-puncher would have to appear on the questionnaire:

Cols 1-5, identical to card 1.
Punch 2 in column 6.

Before sending the completed questionnaires to the card-puncher, you ought to go through them to check that the interviewers have ringed the numbers clearly, and to do any post-coding that is necessary. You will have to decide if there are any cases of 'Inadequate information' and to write in the appropriate code. (You cannot expect the card-puncher to do this for you.) Also, you must tell the card-puncher what to do with questions which have been skipped (i.e. 'Does not apply'); it is usually easiest to leave these columns blank.

Multi-coded answers (see page 82) need special treatment. It is possible to make two holes in the same column of a computer card, but many computers will either reject this or misinterpret it. If there are only two or three possible answers to the question, you can convert them to a single-coded system, as in this example:

Which of the two national newspapers did you read last week?	The Weekly News	1
	The Gazette	2
	Both of these	3
	Neither	4

By having a special code for 'Both of these', you avoid having to punch both 1 and 2.

Alternatively, you can set aside more than one column for the answers to the question. If it was possible for one respondent to give as many as four different answers to the same question, you would set aside four columns on the computer card and punch one answer in each column. But multicoded answers can be awkward to handle in the analysis and it is better to avoid them completely if you are going to use a computer.

As you can see from the illustrations I have given, a computer card contains nothing more than a string of numbers; there is nothing on the card to tell you what the numbers mean. It is essential, therefore, that you keep your own record of this; you have to know that column 7 of card 1 contains the answers to question 1, with code number 1 meaning Yes and 2 meaning No, and so on for all the other information on the cards.

If all the results from a survey can be put onto a small number of cards - less than 2 000 say - you might feed the data cards into the computer every time you want it to do some analysis. If you have many more cards, however, it is more convenient to get the data copied from the cards onto some other data-storing device, such as magnetic tape or disc. (Copying data from cards onto tape or disc is done by machines.) Then, when the computer needs to read the data in order to do some further analysis, it reads it from the tape or disc rather than reading it from the cards again. This is quicker and easier, both for the computer and for you.

Earlier in this section I said that using a computer was like having an assistant. The computer is better than the assistant in carrying out your instructions very quickly, with hardly any mistakes, and without getting tired or bored; but the computer also has weaknesses. Communication is more difficult; a computer is a bit like an assistant who is deaf and dumb, who doesn't understand your language and who cannot read your handwriting. And the computer lacks initiative. You run into both of these problems when you try to tell the computer what you want it to do.

A set of instructions for a computer is known as a 'program'; to get the computer to do any work for you, you have to give it a program. Computers don't understand ordinary English, or any other human language. They have their own languages and you have to use one of these to tell them what to do. Fortunately, a number of program packages have been written specifically for analysing social survey data. One of the best known of these, which is available on many computers, is called SPSS, which stands for 'Statistical Package for the Social Sciences'.

SPSS has been designed so that social researchers can analyse survey data on a computer without having to take a long course in computing. It uses English words and familiar arithmetical symbols. There is a manual which describes it in detail (see Appendix 5); someone with no previous experience of writing computer programs can learn how to use SPSS in a few weeks. I won't attempt to explain SPSS, but, just to give you an idea of what a program looks like, written in SPSS, you might use the following program to obtain the crosstab of Question 5 by village from the poultry survey results:

```
GET FILE        POULTRY

CROSSTABS       VARIABLES = Q5(1,2) VILLAGE (1,9) /
                TABLES = Q5 BY VILLAGE

FINISH
```

You can make use of SPSS only if your computer understands SPSS, of course. If it doesn't, you have to use some other programming language.

How exactly you give the program to the computer depends on the input devices of the computer you are using. I have already described how you can put numerical data onto cards. It is possible also to put programs onto cards and to feed the programs, as well as the data, into a card-reader. Alternatively, you might put your program into the computer using a keyboard that looks like a typewriter, which is attached to the computer.

SPSS, as I have said, has been designed to make computing easier, but even so, you have to be very careful in writing a program. This is because of the second problem that I mentioned - the computer's lack of initiative. If a human assistant is instructed to do some analysis, he has a lot of knowledge about people in general and about this survey in particular, and he uses this knowledge in carrying out the instructions. A computer does not have any of this background knowledge; it just does exactly what it is told. An example might make this clearer.

Suppose you have done a survey of households and, among other things, you have recorded two items of information about each household - the number of children aged 0 to 5 and the number aged 0 to 15. You want to calculate, for each household, the number aged 6 to 15, so you instruct the computer to subtract those aged 0 to 5 from those aged 0 to 15 and to record the answer for each household. (If you have six children aged 0 to 15, and two of them are in the age-group 0 to 5, the other four must be in the age group 6 to 15.) Now, suppose that someone has made a mistake and that one household is recorded as having four children aged 0 to 5 and only two children aged 0 to 15. The computer, unless given specific instructions to the contrary, will subtract four from two, as instructed, and record the answer - 'minus two'. Its arithmetic is correct: 2-4= -2. But the answer is absurd; you can't say, 'We found one household which had minus two children in the age group 6 to 15.' A human assistant could see that this was impossible, but the computer doesn't know it's impossible.

It is easy to make mistakes in writing programs. Often, the mistake will produce results which are obviously wrong, so you can see that you have to correct your program and do the job again. But sometimes the results of a faulty program look reasonable although they are incorrect. For this reason I strongly recommend that you check your programs with a computer specialist.

Before I leave the topic of automatic data-processing, I should just mention another machine called a 'counter-sorter'. This machine takes the same sort of eighty-column cards as a computer's card-reader and it does exactly what its name suggests - it sorts the cards and counts them.

Suppose that the data from each questionnaire of the poultry survey had been transferred onto one computer card, with the answers to question

5 on column 8, and the village code on column 47. In order to get a crosstab of question 5 by village, you would first set the counter-sorter to sort the cards by column 47. You put the cards into the machine, turn it on, and it drops the cards into pockets according to the holes in column 47; all the cards punched 1 on column 47 go into the first pocket, all those punched 2 into the second pocket and so on. You have now divided up the respondents according to their village; all the respondents from the village coded 1 are in the first pocket, all those from the village coded 2 are in the second, and so on.

You take these groups of cards out of the pockets, being careful to keep them separate. Then you set the machine to sort the cards according to column 8. You take the first group of cards (the ones punched 1 on column 47) and run them through the machine again. This time it sorts them according to the holes punched in column 8, and it also counts them. When it's finished this run, it tells you, on a set of dials, how many cards it has dropped into the first pocket, how many into the second and so on. This tells you how many of the people from village number 1 gave the answer coded 1 to question 5, how many gave the answer coded 2 and so on. You repeat this for each group of cards (i.e. for each village) and this gives you the crosstab you wanted. You have sorted the respondents according to their village and you can see the answers they gave to question 5, for each village separately.

This may sound laborious, but the counter-sorter works very fast and you can get a crosstab from a sample of, say, 400 respondents in a couple of minutes. (A clerk using transfer cards would take a couple of hours.)

The counter-sorter is ideal for producing frequency counts and crosstabs from surveys where all the data from each questionnaire can be put onto one card and where there are between 200 and 1 000 respondents. Unfortunately, these machines are not being manufactured any more, and those that still exist are likely to be working badly or out of action for want of spare parts. However, you might have access to one in working order. If so, I recommend that you use it.

Which method to use?

Choosing which method to use will depend on the size of the survey (the number of questions on the questionnaire and the number of respondents) and on the type of analysis you want to carry out. I would put the methods in the following order of increasing sophistication:

1. Counting straight from questionnaires
2. Large sheet of squared paper
3. Transfer cards
4. Edge-clipped transfer cards
5. Counter-sorter
6. Computer

The main advantages of the less sophisticated methods are that you can get tables of results very soon after the end of the fieldwork and that you don't need any special equipment. With the more sophisticated methods you put more time and trouble into converting the data into some special form - transfer cards or computer cards - and you may put yourself at the mercy of machines (which sometimes break down), but when you do start getting tables of results, you can get lots of them quickly. The less sophisticated methods are adequate for the kind of surveys I've been talking about, but you may need a more sophisticated method if you want to do a lot of analysis.

If you want nothing more than frequency counts, and just a few simple crosstabs, the first method is the quickest. If you have fewer than 100 questionnaires and fewer than 60 questions on each, and you will want no more than about 30 crosstabs, the squared paper method is probably the best. For samples of about 100 to 300, it is worth making transfer cards, and it is worth edge-clipping them if you are going to want more than, say, 50 crosstabs, especially if a lot of them will be three-way crosstabs (Chapter 9). If you have access to a reliable counter-sorter and a card-punching service, I suggest you use the counter-sorter method for samples of over 100, so long as you can fit all the data from each questionnaire on one card. Finally, if you don't have a counter-sorter, and if your sample is larger than about 300, or if you are going to want many crosstabs (or other, more sophisticated analysis), it is better to use a computer.

8 Checking survey results

The data you get from a survey may be faulty. It is good policy to carry out some checks on it before proceeding with the analysis.

Response rate *You generally manage to interview only a proportion of the sample you wanted to interview. This proportion is the 'response rate'.*

Comparing the sample with the population *You can sometimes obtain figures on the population to see how representative your sample is.*

Missing data *The data from a survey is often incomplete. This problem needs to be handled carefully.*

Interviewer differences *You should check whether the interviewers have influenced people's answers.*

Range and consistency checks *Checking that code numbers are within the permitted range and that answers are consistent with each other can reveal faults in the data.*

Things can go wrong in a survey. Perhaps the sample was badly drawn, or perhaps the interviewers did not select respondents in the way they were instructed to. Perhaps the respondents misunderstood the questions, or perhaps different interviewers asked the questions in different ways. If things go badly wrong, there is a chance that the results of the survey may be entirely worthless. Suppose that the interviewers realised that they did not have to interview anyone at all; they could drive out of town, have a few days' holiday, fill in the questionnaires themselves, drive back and hand them in. In that case the survey data would be just a pile of rubbish.

Before you spend a lot of effort working on the data, you should carry out some routine checks to see how reliable the results are.

Response rate

In general, you do not expect to interview all the people you want to interview. If you took a sample of rural people over the age of 20, for example, some of them might be very old and too deaf to interview; some might be ill; some might be away from home whenever the interviewers called; and some might refuse to be interviewed. Out of 200 people whom the interviewers visited, perhaps only 186 were actually interviewed. This figure of 186 out of 200 (93%) would be the 'response rate' for this survey.

The response rate will vary a bit from one survey to another, but if the rate for one survey was much higher or lower than usual, you should be suspicious and try to find out why. A response rate of 100% is not necessarily something to be pleased about; it might indicate that the interviewers did not understand the sampling instructions or failed to carry them out.

In Chapter 6, I said that the interviewers should try to obtain some information about the people they cannot interview. This enables you to see how representative the sample is. It is likely that the non-respondents will differ from the respondents in certain ways. For example, in a survey of rural people over the age of ten in Lesotho, we had a response rate of 86%. The interviewers had obtained information on the age and sex of the non-respondents, so we were able to see that about half of the non-respondents were teenage boys. This had come about because teenage boys in Lesotho spend a lot of their time in the fields tending cattle, so they tended to be away from home when the interviewers called. Knowing that the sample was short of teenage boys, we could make allowance for this in presenting the results.

Comparing the sample with the population

The word 'population' has a special meaning in survey research; it means 'the total of people (or things) from which the sample was drawn'. In Chapter 4, I described how to take a sample of 50 co-operative societies out of the total of 700 co-operative societies in the country. In that case, the 700 co-ops were the 'population'. If you took a sample of farmers, your 'population' would be all the farmers in the country.

Sometimes you can obtain figures about your population. If you have interviewed a sample of schoolchildren, you might obtain figures from the Ministry of Education which show how all the schoolchildren in the country (i.e. your population, for this survey) are divided up by sex, by age, by school standard and so on. You can then compare your sample of schoolchildren with the whole population of schoolchildren. To do this, you write the figures for your sample, in percentage form, alongside the figures for the population, also in percentage form.* Perhaps it would look like this:

---

* I explain how to calculate percentages in Chapter 11 (page 142).

		Population	Sample
		%	%
By sex	Male	47	45
	Female	53	55
		%	%
By age	under 8	30	19
	9-12	42	33
	13-16	22	34
	17 or over	6	14
By standard		%	%
	Primary 1-3	37	29
	4 or 5	28	20
	6 or 7	23	32
	Post-primary	12	19

When I put a % sign at the top of a column of figures, I mean that the figures in the column add up to 100%.

By looking at this table, you can see that your sample matches the population quite closely in the ratio of boys to girls; a difference of 2% is nothing to worry about in the type of research we are talking about. When you look at the age table, things are not so good. The interviewers seem to have got too many older children and too few younger ones. Since older children, in general, have had more education than younger ones, the table by educational standard also shows a rather large difference between the sample and the population: the sample contains too many children at the higher levels and too few at the lower levels.

At what point should you start to get worried? This depends partly on the sample size. You do not expect your sample to match the population exactly; on the contrary, you expect to have some discrepancies. The smaller the sample, the larger the discrepancies are likely to be. As a rough guide, with a sample of between 70 and 100 respondents, you should start getting worried if a result from your sample differs from the true figure for the population by more than 10%; with a sample of 100 to 150, you should get worried if it differs by more than 8%; and with a sample of 150 to 200, you should get worried if it differs by more than 6%. (In Chapter 11 and Appendix 1, I explain how these figures are calculated.) If the figures in the previous table had been taken from a sample of more than 100 schoolchildren, the differences between the sample and the population, especially in their ages, would indicate that something had gone wrong in the sampling.

If you find that your sample is not completely representative of the population, you don't necessarily have to throw away the data. It depends on the level of accuracy that you wanted from the survey. To continue the example of the survey of schoolchildren, suppose you undertook it in order to find out, among other things, how many of them could do an arithmetic problem involving four-figure numbers. Children with more education are better at this. So if you found that 18% of the sample could do it, this would be an overestimate of the true figure for all schoolchildren, because your sample has too many children with higher education. Perhaps the true figure is only 10%. Does this matter? You could still say with some confidence, 'Less than a quarter of schoolchildren can do this,' or, 'Only a minority can do this.' Perhaps that was all you wanted to know.

If you do find large differences between the sample and the population, you have to try to find out why this occurred. In an LDTC survey of teenagers, our sample contained more school-attenders than it should have done. It turned out that the interviewers had been unable to reach two of the smaller, less accessible villages that had been selected, and had visited larger, more accessible villages instead. Larger villages are more likely to contain schools, so children in larger villages are more likely to be school attenders, hence the bias of the sample. If you cannot find out why your sample is biased, you have reason to be worried; perhaps the interviewers did other things wrong, in addition to the sampling.

In the example of the schoolchildren, I have used Ministry of Education figures. There are many other sources of figures that you can use for comparing the sample with the population. The Ministry of Agriculture may have statistics on farmers; the Ministry of Employment may have statistics on wage-earners; the Department of Statistics may have reports from the population census and so on.

If you are going to compare your sample with the population in this way, it is a good idea to collect the official statistics on that population before you write the questionnaire; you can then ensure that you collect information in a form that will be comparable with the official statistics. For example, if the census report divides people up into the age-groups 16-20, 21-25, 26-30 and so on, you can arrange your age-groups to match.

Missing data

I mentioned earlier the problem of non-response; out of all the people you wanted to interview, some were not interviewed, so your sample is incomplete. A similar problem can arise when you look at the results for a particular question; some of the respondents who should have answered this question failed to do so. Perhaps they said, 'I don't know,' or perhaps they refused to answer, or perhaps the interviewer omitted the question by mistake, or forgot to write down the answer, or perhaps you have decided that the information is faulty and you have coded it as 'Inadequate information'. Whatever the reason, you have a group of respondents from whom you failed to collect a useable answer to that question. In other words, some of the data is missing.

What you do about the respondents with missing data depends partly on how many there are. If there are very few, you can safely ignore them. For example, if the results for one question from a sample of 143 respondents were, 'Yes 80, No 60, Missing data 3,' it would not make much difference whether these three respondents had answered Yes or No. If all three had answered Yes, the result would have been 'Yes 58%, No 42%.' If all three answered No, the result would have been 'Yes 56%, No 44%.' The difference is not worth worrying about. You could simply exclude them and present the result as 'Yes 57%, No 43%' (i.e. 80 out of 140, and 60 out of 140).

But if the data is missing for a large proportion of the respondents, you have to be more careful. Suppose you had the following figures from a survey of 91 correspondence students:

In the last week, about how many hours have you spent studying at home?	10 hours or more	23
	More than 5 but less than 10	32
	5 hours or less	14
	Missing data	22
	Total	91

If you excluded the 22 'missing' students from the total (91-22=69), you would calculate that 33% had studied for ten hours or more (23 out of 69). But what if all 22 of the 'missing' students had actually studied for less than ten hours? The true result would be that only 25% of the students (23 out of 91) had studied for ten hours or more. Worse still, what if all 22 had studied for more than ten hours? The true result would then be that 60% of the students (55 out of 91) had studied for ten hours or more.

The question you have to face is whether the 'missing data' group are likely to be much the same as the rest or very different from the rest. If they are likely to be much the same, you exclude them from the total. What you are saying, in effect, is this: 'Out of the 69 who answered the question, 33% said "ten hours or more". We think that, out of the 22 whose data is missing, the same proportion (about 33%) would have said "ten hours or more".' But if the 'missing data' group are likely to be very different from the rest, you should not do this.

Sometimes you can make a good guess if you know why you failed to get information from these people. In this example, if the reason for the 22 missing answers was simply that 22 of the questionnaires had a page missing, there would be no reason to suppose that these 22 students would have answered that question very differently from the other 69; in that case you would take 69 as the total. But suppose you had the following results from a survey of farmers:

How many cattle do you own?	None	24
	1-10	62
	11-20	10
	More than 20	3
	Missing data	43
	Total	142

If you excluded the 'missing data' group, you would calculate that 3% of the farmers had more than twenty cattle (3 out of 99). But perhaps the 43 'missing' farmers had all refused to answer the question because they were rich men who wanted to conceal their wealth from the interviewer. If all 43 were in fact rich cattle-owners, the result should be that 32% owned more than twenty cattle (46 out of 142), not 3% (3 out of 99).

In such a case, the safest thing to do would be to discard the results for that question, but you might not want to do that. Perhaps the main purpose of the survey was to compare the rich farmers with poor farmers, so this question is crucial to the whole survey. In that case, you should compare the 43 'missing' farmers with the other 99 on other questions in the questionnaire. For example, if you found that these 43 also had large houses and held positions of authority in their villages, you would know that they were, almost certainly, richer than the rest. By contrast, if you found that their other answers did not differ much from those of the other respondents, this would suggest that your suspicions were unfounded; the 43 'missing' farmers were probably not concealing their wealth from the interviewer; there must be some other reason why their answers are missing. If you did present a result for this question in your report on the survey, you should certainly draw the reader's attention to the high proportion of missing data and explain that the survey result might be very wide of the mark.

Interviewer differences

Interviewers are human. If the wording of a question seems inappropriate for a particular respondent, the interviewer will, naturally, tend to modify it. If a question seems too long, he will be inclined to shorten it. If a respondent seems not to understand the question, the interviewer will rephrase it or explain it in his own words. If an interviewer doesn't like asking a question because it causes embarrassment, he will be tempted to miss it out completely.

The trouble is that each interviewer modifies the questions in his own way. It is as if each interviewer was using a slightly different questionnaire. If one interviewer is consistently asking a slightly different question from another interviewer, he will tend to get a different pattern of results.

A further problem is when people give answers that the interviewer does not know how to record. This can be the result of poor questionnaire design. Take an example that I mentioned in Chapter 5:

What do you think is the ideal number of children for a couple to have today?

Three or fewer ☐
Four or more ☐
Don't know ☐

What does the interviewer do if someone answers, 'God will decide'? One interviewer might think this means that the respondent disapproves of contraception and therefore, by implication, that he is in favour of large families; so the interviewer marks 'Four or more'. Another interviewer might think that the respondent is avoiding the question, so he marks 'Don't know'. Yet another interviewer might refuse to accept this answer, since it is not catered for on the questionnaire, and insist that the respondent picks a number. Each of these three interviewers might in fact receive this answer the same number of times, but the questionnaires they hand in will contain three different sets of results for this question.

As a consequence of all this, there are likely to be a few questions on a questionnaire for which different interviewers get completely different results. I mentioned in Chapter 6 that it was a good idea to arrange the interviewing so that all the interviewers visited the same places and interviewed the same sorts of people. If the interviewing has been done in this way, the results from one interviewer should be similar to those from another interviewer, so you can compare the results obtained by different interviewers. If there are questions for which they have got very different results, they must have been handling those questions differently. (If the interviewing has not been arranged in this way, you cannot make this check. If one interviewer did one village and another did another village, any differences between their results could be due to a difference between the villages. Or if one interviewer did men and another did women, the differences could be due to the sex of the respondents.)

To check for interviewer differences, you put the results obtained by different interviewers alongside each other. The following table shows how this might look, for a question with only two possible answers, from a survey in which four interviewers interviewed fifty people each. The table shows, first, the actual numbers of respondents and then the same results expressed as percentages:

	Judy	Janet	Josephine	Jill	Total
Yes	20	22	19	35	96
No	30	28	31	15	104
Total	50	50	50	50	200
	%	%	%	%	%
Yes	40	44	38	70	48
No	60	56	62	30	52

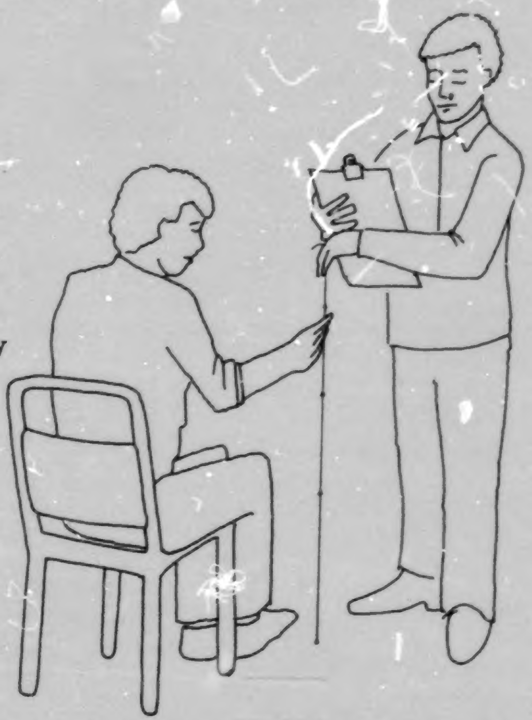
You would expect the results of these four interviewers to be roughly the same, but not exactly the same. The small differences between the results obtained by Judy, Janet and Josephine are nothing to worry about. Jill's results, however, are more seriously out of line, suggesting that Jill was handling this question in a different way from the others.

The same question arises here as with defects in the sample: how big do the differences have to be before you start getting worried? Again, it depends partly on the number of respondents that each interviewer interviews. If each interviewer has interviewed only a few people, you would expect quite large discrepancies between their results. In fact, if the interviewers have interviewed fewer than twenty people each, it is probably not worthwhile to compare the results of different interviewers. As the number gets larger, the discrepancies ought to get smaller. As a rough guide, if each interviewer has interviewed more than forty respondents, you should begin to suspect that something has gone wrong if one interviewer's results differ from another's by more than 25%, especially if it is the same interviewer who is out of line on several questions.

And again, it depends on the level of accuracy you want. Perhaps you wanted to know whether the proportion of people listening to a radio programme was about a quarter of the population, or about half, or about three-quarters. If two interviewers get different results because of recording a small proportion of their answers differently - say one gets '17%' and the other '32%' - this does not matter too much since, whichever is right, you have an answer to your question, namely 'about a quarter'.

Sometimes, the discrepancies are so large that you really do not know what you have found. If one interviewer's results say that 20% of the respondents had listened to this programme, whereas another interviewer's results say that 70% had, then something has gone wrong. Occasionally, you can find out what went wrong by discussing it with the interviewers. The best way to approach this is, first, to ask each interviewer on his own to demonstrate to you how he asked those questions, and how he recorded various possible answers, and then to bring all the interviewers together for a joint discussion of the problems.

For example, in one survey at LDTC we wanted to know what proportion of people knew how long a metre was. We asked them to hold their hand one metre above the ground. The interviewer then took out a measuring string that we had made - a length of string divided into sections painted different colours. What he had to do was hold up the string so that the weight on the end touched the floor, and then note down which colour the respondent had his hand next to, as in the illustration. If the respondent's hand was next to the orange section, he was close to one metre; if he was by the blue section, he was too low, and if by the red section, he was too high. The interviewers got completely different results. When they were asked to demonstrate exactly how they had asked this question, it became clear that one of them was not sure which end of the string was the top end; sometimes he had used it correctly and sometimes he had held it upside down.



Unfortunately, it is often impossible to find out why the interviewers got different results. In these cases, one has no alternative but to discard the data for those questions. This is painful, but if you don't know whether the '20%' of the first interviewer is the right answer or the '70%' of the second interviewer (or the '45%' you get by combining them), you really have not found out anything at all.

Range and consistency checks

The data from any survey will contain some mistakes; an interviewer might have recorded an answer wrongly, or the answer might have been transferred incorrectly onto a transfer card, or punched incorrectly onto a computer card. (These terms were explained in Chapter 7.) One way to find a few of these mistakes is to check that the code numbers that have been recorded or punched are at least possible. For example, the possible codes for a question might be 1 for 'Yes', 2 for 'No', 9 for 'Don't know', 0 for 'Inadequate information' and blank for 'Does not apply'. If any other code number has been recorded or punched for this question, such as a 3 or a 5, it must be a mistake. This is known as a 'range check'. If the data has been fed into a computer, it should be possible for a computer specialist to write a program, or to make use of a ready-made one, which will perform this range check automatically.

Another way of finding mistakes is to see if the answers that have been recorded for each respondent are internally consistent. On questions of opinion, people are often inconsistent, so you cannot always expect that all the answers which someone gives to a questionnaire will hang together perfectly. On the other hand, you would normally expect a person's answers to be consistent on matters of fact. If a woman says that she has no children, you would not expect her to say, later in the interview, that she has five children. If you did find this sort of self-contradiction in a person's answers, you would suspect that something had gone wrong.

To do a consistency check, you select two questions such that the answers to one question should be related to the answers to the other. For example, the answers that children give to the question, 'What did you do yesterday?' ought to be related to their answers to 'Are you attending school?' if the survey is conducted during the school term. If anyone said he was not attending school, but that he had attended school the day before the interview, you would suspect that something was wrong. You could test this by getting a cross-tabulation, in this form:

		Are you attending school?		
		Yes	No	Other
What did you do yesterday?	Went to school		*	
	Other			

The box marked with an asterisk is the important one. There ought to be no respondents there at all.

If you thought, in advance of doing this survey, that there was a serious possibility of respondents giving false answers, you could make a point of writing pairs of questions like this into the questionnaire, putting them into different parts of the questionnaire, purposely to detect whether people gave inconsistent answers.

If you find that some of the data is inconsistent on matters of fact, all you know is that something has gone wrong. It does not follow necessarily that respondents were giving false answers. That, of course, is one possibility, but there are others; perhaps the respondents did not understand the questions, perhaps the translation was poor, perhaps the interviewers recorded the answers wrongly, or perhaps these answers were coded wrongly, or punched wrongly.

Occasionally, it is possible to check the answers of different respondents against each other. In an LDTC survey of rural teenagers, we interviewed both the teenagers and their parents. We could ask the teenagers some questions about themselves (e.g. 'What standard are you in at school?') and we could ask their parents the same question about the teenagers (e.g. 'What standard is he in at school?') Then we could cross-tabulate the answers of the teenagers against the answers of their parents.

To sum up these checks, the response rate may warn you that the interviewers did not select respondents in the way they were meant to. Comparing the sample with the population may show that your sample is unrepresentative. Looking at interviewer differences may show that the interviewing was not carried out correctly in every detail. Range and consistency checks, along with an inspection of missing data, may reveal that respondents were misunderstanding questions, or avoiding them, or that errors were made in recording or transferring the data.

Having made some assessment of the reliability of the data, having discarded data that is clearly faulty and having cleaned up as many as possible of the mistakes which have crept in, you are ready to proceed to the next stage - analysing the results.

It takes some self-discipline to carry out these checks. It's like going out of your way to hear some bad news. By the time you reach this stage of a survey, you have already put a lot of effort into it and, naturally, you would rather not know if something has gone wrong. A lazy researcher will skip the checks and go straight on to analyse the data, simply assuming that the data is fine. The risk he runs is that some of his results - perhaps all of them - may be complete rubbish.

9 Analysing and reporting survey results

This chapter describes some straightforward techniques for making sense of the figures you get from a survey.

Reweighting This is a procedure you can sometimes use to correct for a systematic bias in the sample.

Cross-tabulations You divide the respondents into separate groups and compare their answers.

Hidden factors A pattern of results may have more than one explanation. The first one you think of may not be the correct one.

Diagrams: pie-charts and bar-charts Two ways of presenting results in diagram form.

Writing the report Keep the report as short, clear and readable as possible.

Completing the course A survey is a big job. Try to see your way through to the end of it before you decide to begin.

The analysis of a small survey can be fairly straightforward. Indeed, unless you have access to a computer, it has to be straightforward. The simplest questions to answer are the 'How many?' type, e.g. 'How many households own poultry?' You just count up the answers to each question and present them as percentages. You are showing how frequently each answer was given, so these results are sometimes called 'frequency distributions'. (I introduced the term 'frequency count' at the beginning of Chapter 7.) The main problems with this are deciding what to do with 'missing' cases and selecting the most appropriate base for each set of percentages. I discussed the problem of missing cases in the previous chapter and I deal with percentages in Chapter 11.

Occasionally it happens that your checks on the data, such as the ones I recommended in Chapter 8, reveal some fault which makes you think that your straight survey results are probably wide of the mark. If, as is often the case, there is one particular result that is especially important, it may be worthwhile attempting to correct for the fault by a procedure that is known as 'reweighting the data'.

Reweighting

Suppose you want to find out what proportion of the adult population can read, and you have given a reading test to a sample of 240 people. In checking the results, you have discovered that there was some defect in the sampling. You know from official statistics that only 40% of the adult population have had any schooling, yet you have found that 55% of your sample have had some schooling. This is serious because, as you expected and as the results show, there is a close connection between schooling and the ability to read. Let's say your results on this were as follows (first as raw figures, then as percentages):

	Had had some schooling	Had had no schooling	Total
Could read	117	8	125
Could not read	21	94	115
Totals	138	102	240

	Had had some schooling	Had had no schooling	Total
	%	%	%
Could read	85	8	52
Could not read	15	92	48
Base totals	(138)	(102)	(240)

Your result - that 52% can read - is likely to be an overestimate of the true figure because your sample contains too many educated people. You might say that your sample was too heavy on the side of educated people. To correct for this, you can adjust the weights, hence the term 'reweighting'.

You begin by pretending that you had interviewed 100 people and that you had the correct balance of educated and uneducated people. Call this your 'imaginary sample'. The bottom line of the table would look like this:

	Had had some schooling	Had had no schooling	Total
Could read Could not read			
Totals	40	60	100

You know from the survey that 85% of those who had had some schooling could read. So you can guess that, of the 40 people in your imaginary sample who had had some schooling, 85% could read. 85% of 40 is 34. So the first column in the table would look like this:

	Had had some schooling	Had had no schooling	Total
Could read	34		
Could not read	6		
Totals	40	60	100

Similarly, you know that 8% of those who had had no schooling could read. 8% of 60 is 5. So the second column would look like this:

	Had had some schooling	Had had no schooling	Total
Could read	34	5	
Could not read	6	55	
Totals	40	60	100

By adding up these figures, you can fill in the right-hand (total) column of the table:

	Had had some schooling	Had had no schooling	Total
Could read	34	5	39
Could not read	6	55	61
Totals	40	60	100

So, 39% of your imaginary sample could read. This is the same as saying that, if you had had the correct proportion of educated people in the sample (i.e. 40% instead of 55%), you estimate that you would have found that 39% of the total sample could read. You have corrected the weights.

I have to admit that this procedure is unsatisfactory. In particular you have to assume that the 138 educated people in the sample were representative of all educated people, and that the 102 uneducated people in the sample were representative of all uneducated people. This may not be true. If the sample was defective in one way (i.e. the proportions of educated to uneducated people), it may be defective in other ways too. Certainly you should not reweight results if the sample is extremely defective. If your sample had contained only 24 uneducated people (10%), the result of reweighting would be so unreliable as to be quite useless.

It is far better to take some care over getting a good sample in the first place than to tinker with the figures at the end. However, accidents do happen and you might prefer to salvage something from a survey rather than to discard the results altogether. In this example, you might argue that, though 39% is not likely to be completely accurate, it's likely to be closer to the truth than the original 52%.

Cross-tabulations

After working on the frequencies for some time, so that you have a good idea of the basic results of the survey, you then want to proceed to the second sort of question, which takes the form 'How does this group of people differ from that group?' In a family planning survey, you might want to compare the answers of men and women; in a literacy survey, you might want to compare the more educated with the less educated. In other surveys, you might want to compare older people with younger people, rural mothers with urban mothers, students of maths with students of English, households who have radios with households who don't, and so on. To make these comparisons, you arrange your results in a cross-tabulation, or 'crosstab' for short.

I need to introduce here the term 'variable'. (I explain it in more detail in Chapter 11, in the section headed 'Category variables and measurement variables'.) A variable is some characteristic that varies from one person to another. Age, for example, varies from one person to another, so age is a variable. In survey analysis, the term 'variable' is often used to refer to any of the items of information that you have on each respondent. From a survey of poultry-keeping, you might have details of the sex, age, marital status, and educational level of each respondent, and also the number of birds he owns, the type of food he gives them, the people he seeks advice from, his opinions about improved breeds and so on, and each of these things is a variable.*

In Chapter 7 I explained how to use transfer cards to produce a crosstab. Briefly, you first sort the cards into different piles and then you count the answers of each pile separately. If you want to compare the answers of men and women to question 29, you sort the cards into one pile of men and another of women, then you count their answers (perhaps by the tally-mark method) to question 29. Here you are considering two variables - 'sex' and 'answers to question 29'. You might describe this crosstab as 'question 29 by sex'. The variable that you sort the cards on (sex, in the example) is called the 'independent variable', but I will call it the 'sorting variable' as I think this is clearer. The other variable ('answers to question 29') is called the 'dependent variable', but I will call it the 'counting variable'.

You have to think carefully in order to get the variables the right way round in a crosstab. Say you want to know whether teenagers are more literate than adults. The raw figures might look like this:

---

* Readers who know some statistics may have come across a distinction between 'variables' and 'attributes'. I use the term 'variable' for both, but I will distinguish in Chapter 11 between category variables (which some people call 'attributes') and measurement variables.

	Teenagers (age 13-19)	Adults (age 20 or over)	Total
Could read	27	98	125
Could not read	13	102	115
Total	40	200	240

For each of the cells of a crosstab there are three percentages that you might calculate - out of the row total, the column total and the grand total. For example, the 27 teenagers who could read form 22% of the row total (all 125 people who could read), or 68% of the column total (all 40 teenagers), or 11% of the grand total (all 240 people in the table). Generally it is clearest to arrange a crosstab so that the sorting variable forms the columns, the counting variable forms the rows, and only the column percentages are shown. To see whether teenagers are more literate than adults, you use age as the column variable and literacy as the row variable, and you present the table like this:

		Sorting variable	
		Teenagers	Adults
		%	%
Counting variable	Could read	68	49
	Could not read	32	51
	Base totals	(40)	(200)

This shows clearly that the teenagers in the sample were more literate than the adults. If you used literacy as the sorting variable and age as the counting variable, the table would look like this:

		Sorting variable	
		Could read	Could not read
		%	%
Counting variable	Teenagers	22	11
	Adults	78	89
	Base totals	(125)	(115)

This table reports exactly the same raw results as the first table, but it presents the results so that they show something different. The first table shows how many teenagers can read, compared to adults. The second table shows how many readers are teenagers, compared to non-readers.

A variable might appear as the counting variable in some tables, but as the sorting variable in others. Say you have interviewed people about a road safety campaign. You might first use your results to see whether men are more likely than women to be drivers. In this table, sex is the sorting variable and driving is the counting variable, so you would prepare the table like this:

	Men %	Women %
Driver		
Non-driver		
Base totals		

Elsewhere in the report you might want to see whether drivers were more likely than non-drivers to have seen the road safety posters. In this table, driving would be the sorting variable, like this:

	Drivers %	Non-drivers %
Had seen poster		
Had not seen poster		
Base totals		

If in doubt about which variable is the sorting one and which is the counting one, fill the blanks in this sentence:

I am comparing .... with ...., with respect to ....
e.g. I am comparing teenagers with adults, with respect to literacy.

$\underbrace{\text{teenagers with adults,}}_{\text{The variable you insert here is the sorting variable.}}$ $\underbrace{\text{literacy.}}_{\text{The variable you insert here is the counting variable.}}$

The examples of crosstabs I have given here have all had two columns and two rows. These are called 'two-by-two' tables. Some variables, of course, can have more than two categories, so you can have tables with more than two rows and columns. This one, for example, which you might describe as 'schooling by age group', is a 'three-by-five' table:

	21-30 %	31-40 %	41-50 %	51-60 %	61 or over %
Had no schooling	13	17	24	33	53
Left school at St.3 or below	37	47	47	50	38
Left school at St.4 or above	50	36	29	17	9
Base totals	(40)	(36)	(34)	(31)	(32)

Hidden factors

Crosstabs can be misleading. You have to be careful when you draw conclusions from them. Say you got the following results, in Lesotho, from a survey on family planning:

Table A		Men %	Women %
Have you heard of family planning?	Yes	35	47
	No	65	53
	Base totals	(198)	(182)

It appears that sex is an important factor which influences whether or not people get to hear of family planning - women are more likely than men to have heard of it. One might speculate that this is because women take more interest in family matters than men, or because women attend clinics (with their infants) more than men do, so they are more likely to see posters or to hear lectures about family planning. But, in fact, this speculation would be premature. A further crosstab might reveal the following:

Table B		Had had some schooling %	Had had no schooling %
Have you heard of family planning?	Yes	60	16
	No	40	84
	Base totals	(207)	(153)

It appears that school education is also an important factor - those with some schooling are more likely to have heard of family planning than those with no schooling. Now, it happens that, in Lesotho, girls tend to receive more schooling than boys. Let's say that this is reflected in this sample: 70% of the women in the sample have had some schooling as against only 45% of the men. If school education is an important factor, you would expect more women than men to have heard of family planning, simply because more women than men have had a school education. Does this explain the apparent difference between the sexes that we found in the first table?

To test this, you need a three-way crosstab. In the previous section all the examples I gave were two-way crosstabs. A two-way crosstab is one in which you have two variables - a sorting variable and a counting variable. (Do not confuse a 'two-way' table with a 'two-by-two' table; 'two-way' means there are two variables, while 'two-by-two' refers to the number of rows and columns.) A three-way crosstab is one in which you have three variables - two sorting variables and one counting variable.

In this example, you would first sort the cards into two piles, one for 'Had had some schooling', the other for 'Had had no schooling'. So your first sorting variable is 'Education'. Then you take the first pile and sort it into men and women, i.e. your second sorting variable is 'Sex'. Then you count the Yes/No answers of the men to the question 'Have you heard of family planning?' And then you count the Yes/No answers of the women, i.e.

the counting variable is 'Whether heard of family planning'. This gives you a crosstab like this:

Table C1		Had had some schooling	
		Men	Women
		%	%
Have you heard of family planning?	Yes	60	60
	No	40	40
	Base totals	(80)	(127)

But you have not finished yet. Now you go back to the pile of cards for 'Had had no schooling'. Again, you divide men from women and then count up their answers to 'Have you heard of family planning?' This gives you a second crosstab. If you put it below the first crosstab, the two together look like this:

Table C1		Had had some schooling	
		Men	Women
		%	%
Have you heard of family planning?	Yes	60	60
	No	40	40
	Base totals	(80)	(127)
Table C2		Had had no schooling	
		Men	Women
		%	%
Have you heard of family planning?	Yes	14	18
	No	86	82
	Base totals	(98)	(55)

This is a three-way crosstab. You could call it 'Whether heard of family planning by sex by education'.

The table shows that sex is not really an important factor in itself. Table C1 shows that educated men are just as likely to have heard of family planning as educated women. Table C2 shows that uneducated women are not more likely to have heard of family planning than uneducated men. (The difference of 4% can be disregarded, for reasons I explain in Chapter 11 in the section on 'statistical significance'.)

Table A in this example appeared to show that sex was a factor that influenced people's awareness of family planning. It was tempting to conclude that women were more aware of family planning just because they were women. But there was a hidden factor, namely school education. What Table A really showed, indirectly, was just that the women had had more schooling than the men.

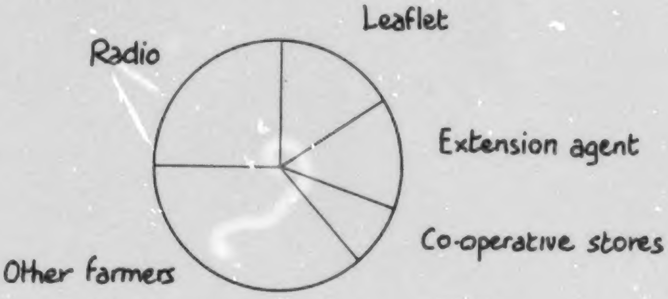
Hidden factors are the bugbear of science in general and social science in particular. If ever you want to claim that factor A is responsible for a particular result (e.g. a person's sex affects his awareness of family planning), you should always pause to think whether there might be a hidden factor B, which is linked to factor A and which is the one that is really responsible for the result. If you find that farmers who have read your pamphlet are better informed about fertilisers than farmers who have not read it, you should not immediately conclude that it is because of your pamphlet. Perhaps those same farmers who have read your pamphlet have also attended lectures and listened to radio programmes about fertilisers; or perhaps literate farmers are more informed than illiterate ones anyway.

Diagrams: pie-charts and bar-charts

Diagrams can sometimes make results clearer, especially for readers who don't like columns of figures. A simple way to represent a frequency distribution is in the form of a pie-chart. In a pie-chart, you divide up a circle, like cutting pieces of a round pie, so that the size of the slices corresponds to the proportions in the distribution. For example, the pie-chart on the right corresponds to the table on the left:

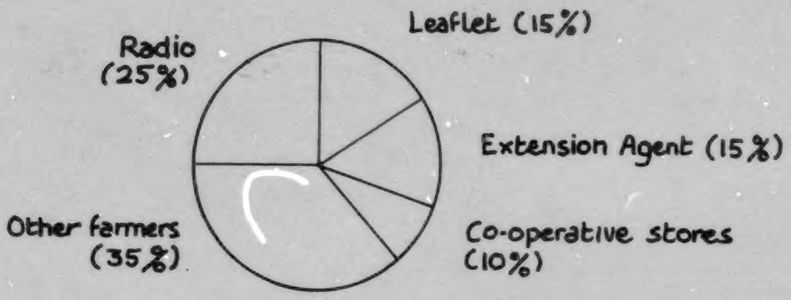
Farmers who had heard about the new seed had first heard about it from:

	%
Radio	25
Leaflet	15
Extension agent	15
Co-operative stores	10
Other farmers	35
Base total	(220)



The angles formed by the slices at the centre of the pie add up to 360°; in other words, 360° represents 100%. To calculate the angle which will represent, say, 15%, you work out $360 \times \frac{15}{100}$, which comes to 54°. To measure angles when drawing a pie-chart, you need a protractor (a small semicircle of clear plastic with angles marked on it).

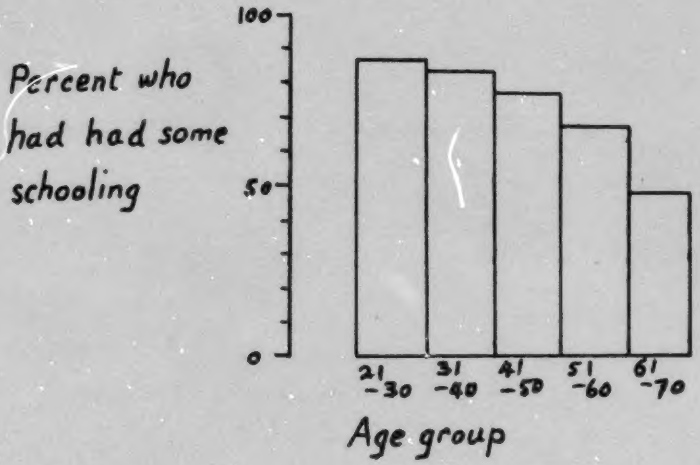
With pie-charts, as with all diagrams, you should present the actual figures as well as the diagram. You might do this by presenting both the table and the diagram, as above, or you might include the figures in the diagram, like this:



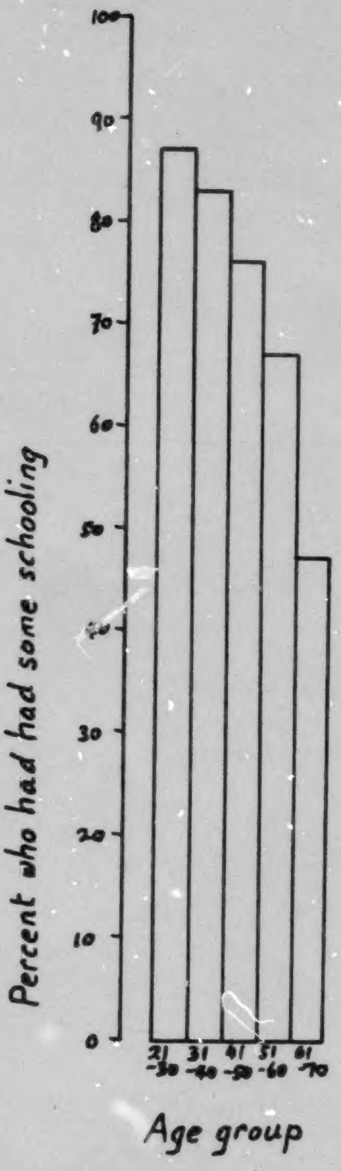
Total number of farmers who had heard about the new seed (=100 %) 220

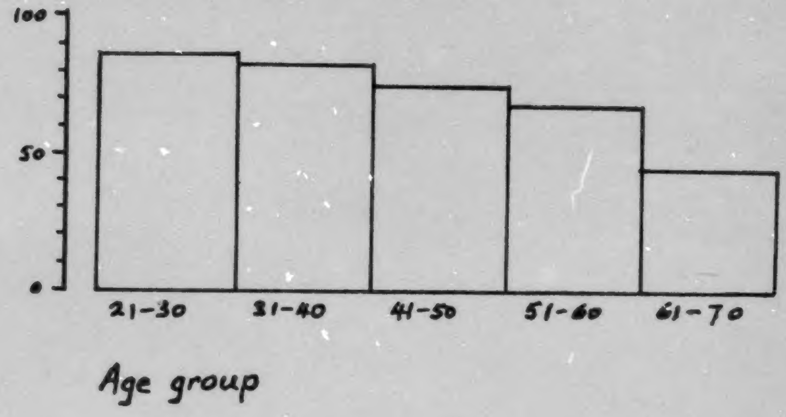
To represent a crosstab you need a different sort of diagram, called a bar-chart or 'histogram'. Here is an example of a bar-chart along with the crosstab that it represents:

	Age 21-30 %	31-40 %	41-50 %	51-60 %	61-70 %
Had no schooling	13	17	24	33	53
Had some schooling	87	83	76	67	47
Base totals	(40)	(36)	(34)	(30)	(32)

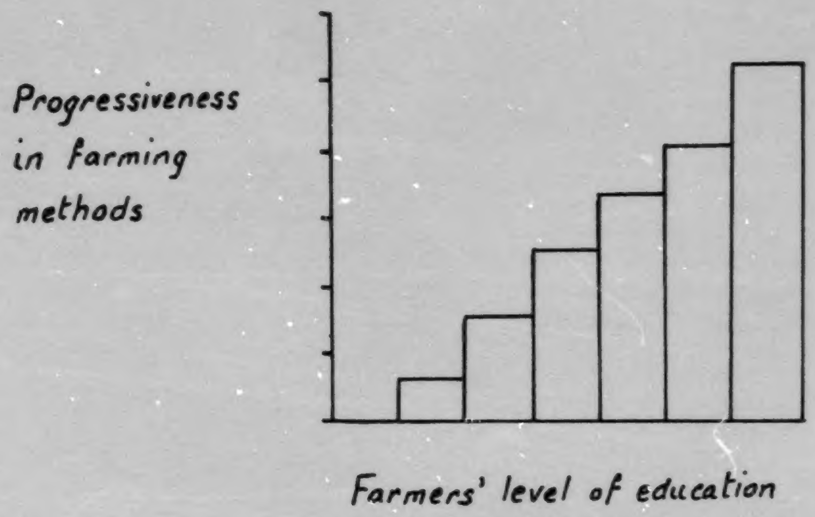


When designing a bar-chart you should try to make it approximately square. If the vertical distances are either unnecessarily long or unreasonably short in comparison with the horizontal line, the effect is misleading. The next two diagrams present the same information as the one above, but the first makes the differences between the age groups seem large while the second makes them seem small.


Percent who had had some schooling
100
90
80
70
60
50
40
30
20
10
0
21-30
31-40
41-50
51-60
61-70
Age group

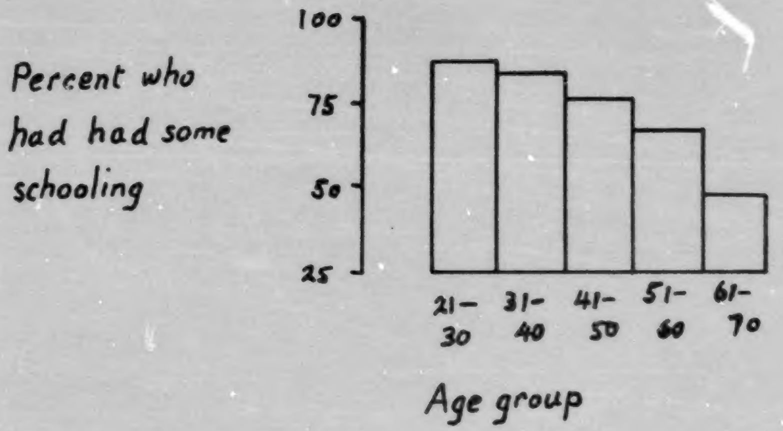

100
50
0
21-30
31-40
41-50
51-60
61-70
Age group

There are two faults that one sees quite often in bar charts. One of these is to present the diagram without fully labelling the scales. What does this bar chart show, for example?



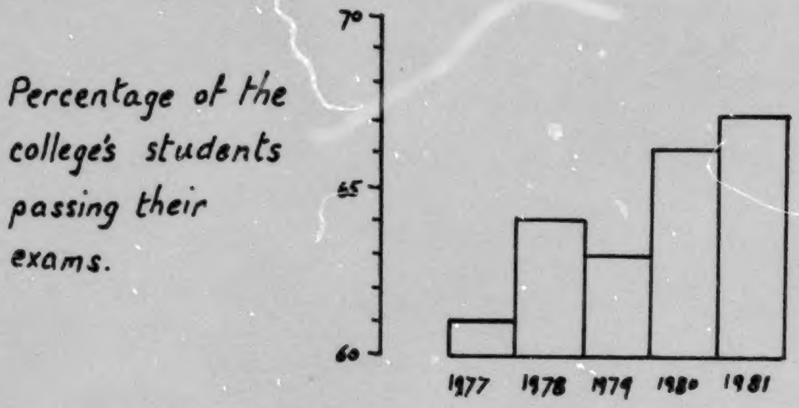
At first glance it appears to show that farmers with a higher level of education are more progressive, but the levels of education are not marked, nor is there any indication of how 'progressiveness' was measured. Without these items of information, the diagram is virtually meaningless.

The second common fault is to present only the upper part of the bar chart. This exaggerates the differences between the groups. If you did this with the earlier diagram on schooling and age group, for example, it might look like this:

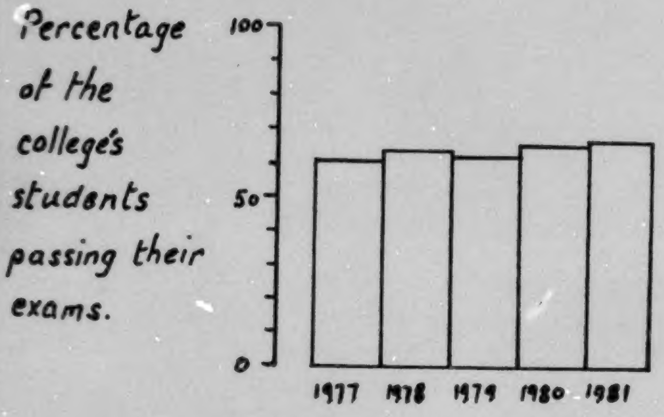


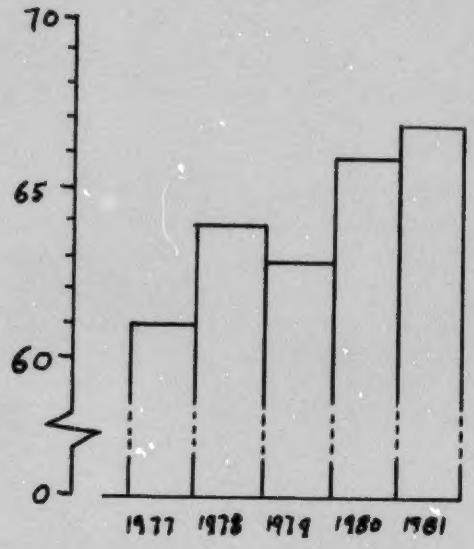
By setting the horizontal axis level with 25% rather than 0%, this diagram makes the differences between the age groups seem larger than they really are.

Occasionally people purposely incorporate this fault into a bar chart with the full intention of misleading the reader, as in this example:



The college's success rate seems to be improving by leaps and bounds, but the improvement is really quite modest, from 61% in 1977 to 67% in 1981. The diagram ought to be presented in one of these two ways:





If you use diagrams in a report, use them sparingly, to illustrate a few important findings. If a reader finds just two or three diagrams in a report, there is some chance that he will pay attention to them, but if he is faced with pages of diagrams, he will skip them.

Writing the report

The purpose of the survey report is to communicate the results of the survey to people who have an interest in them in a way that they can understand. This might seem obvious and yet many research reports are written in such a stilted 'scientific' style and filled with so much technical detail that it is actually difficult to see what the results were.

You usually begin with a short explanation of why the survey was conducted, outlining the policy questions that the research was intended to illustrate. Then you describe how the survey was done. You should mention at least the following points:

Sampling	how it was done, the number you wanted to interview
Response	the number you did interview, the response rate, characteristics of non-respondents, reasons for non-response
Interviewing	how it was done, when it was done
Checks on the data	your assessment of the reliability of the results, particularly any sample bias
Confidence limits	
Significance tests	which tests you used and what level of significance you adopted (more on this in Chapter 11)

In the main report you should be fairly brief in describing how the survey was done. Most of your readers will not be interested in the technical details, and many will not even understand them. You should put detailed technicalities into an appendix.

The main part of the report is the presentation of the results. Many people find it difficult to take in long strings of results, especially when the results are in figures (numbers). You should select the most important findings and make sure that they stand out clearly in the report, so that a reader will be sure to take in those points, even if he misses everything else. You should break up the results into the different topics and deal with each topic separately. Use the text to guide the reader about what he is to look for in the tables, in this sort of way - 'Whether or not a person has heard of family planning depends very much on the amount of formal education he has received. This is shown in the following table ....'

The tables themselves should be as clear and as simple as possible. Many readers are inclined to ignore tables of figures completely and they will certainly do so if the tables are complicated. At the same time, the tables should not be misleading. If the table presents people's answers to a rather complicated question, write out the question in full as part of the table so that the reader knows exactly what he is looking at. Tables of percentages should contain the base totals.* If the base total is restricted to only a part of the sample (e.g. 'poultry owners' or 'women who had heard of family planning') this should be written out in the table.

There are two common errors that I have noticed people make in interpreting tables of figures; it is worthwhile taking special care to prevent readers from making them. The first is a mistake about base totals. The result might be 'Of those people who could read, 95% had learnt at school.' This is interpreted as '95% of people could read and they learnt to do so at school.' The second is seeing differences when really they should see similarities. If a survey found that 20% of households in the northern district had radios, 17% of households in the eastern district and 15% in the southern district, the important thing would not be 'Radio ownership is higher in the northern district,' but rather, 'Radio ownership is at much the same level throughout the country.' This depends on the context; a difference of 5% might be very important in some studies. But, in general, people tend to exaggerate the importance of small differences.

---

* See Chapter 11 (page 142) for an explanation of the term 'base total'.

It is possible for a report to end at this point, after the presentation of the results. However, the research I have been talking about is intended to guide policy in some way; someone is supposed to take some action on the basis of the results. It is natural, then, for the report to have a further section - conclusions and recommendations. This raises special problems. These problems, however, are not technical ones connected with survey methods, but rather 'political' or organisational ones, arising out of the relation between research and the rest of the distance-teaching institution. For that reason, I will defer them to the final chapter.

Keep reports short. A large survey might deserve a long report, but you generally can't expect readers to take in more than five or six points, and you can probably put these in a report of five or ten pages. After all, a report is useless if no-one reads it, and people are more likely to read a short report than a long one. If the report has to be long, present the main points in a summary at the beginning.

Completing the course

If you have read to here from Chapter 4 you might have formed the impression that doing a social survey is a bit like running the marathon. It is. Like marathon runners, many people who begin a survey fail to complete the course. I don't know if anyone has done a survey of surveys, but I would guess that between one third and two thirds of all surveys that get started never get written up in a final report.

One reason is that people do not foresee all the stages they will have to go through; they might administer a questionnaire to 200 people and then find that they do not know what to do with this large pile of questionnaires. Another reason is that they allow part of the survey to get out of proportion with the rest. Experts are sometimes to blame here for pronouncing on how a part of the work, such as the sampling, ought to be done, without appreciating the limited resources of the agency doing the survey. It is no good attempting a survey of 1 000 households if you haven't the time or the staff to train the interviewers properly, or if you haven't the equipment to process the large amount of data that will result.

Try to see your way through to the end of a survey before you begin it. Of course you cannot plan every step in detail, but you should have some idea of how you are going to tackle each stage, including the later stages of data processing and analysis. Try to be realistic about the time it will take. At LDTC we found we could do a quick survey - say a questionnaire of 30 questions to 100 rural adults - in about three months, but a larger survey - say 100 questions to 300 rural adults, carefully sampled - took 18 months to two years, from drafting the pilot questionnaire to printing the final report.

Match the scale and sophistication of the survey to the requirements of the educators who have commissioned it. Large-scale surveys that are carried out by governments or big market research agencies are intended to provide results that are accurate to within a few per cent, and these are the surveys generally described in survey textbooks; but educators will often be content with results at a very rough level of accuracy - 'About a fifth', 'Roughly half', 'Almost all' and so on. And, finally, do not ignore other research methods. Make full use of published statistics and of records collected by other agencies. Consider whether observation or discussion will suffice. If, for example, you wanted some idea of the range of opinions on some topic but didn't need to know the proportions of people holding this or that opinion, a few group discussions might be better than a survey. You should only set out on the marathon of a survey if you are fairly confident you can complete the course in reasonable time and if you think that that much effort is justified.

10 Experiments

Experiments can be carried out in distance teaching but they have to be designed carefully to produce useful results.

Internal validity *A poorly designed experiment may not really prove what it seems to prove.*

Some technical terms *'Hypothesis', 'control group', 'pre-test and post-test', 'experimental subjects'.*

Comparing like with like *Assigning people at random into groups gives you groups that are approximately equivalent.*

External validity *You cannot always assume that what happens in an experiment will happen in real life.*

Balancing internal and external validity *Being able to generalise from the results of an experiment is more important in practical research than in academic research.*

Real-life experiments, action research and pilot projects *Experiments are most useful when conducted as part of real-life operations.*

This chapter does not have a section specifically on the analysis and reporting of experiments. Much of Chapter 9 is relevant to experiments as well as to surveys and the necessary statistical methods are explained in Chapter 11 and Appendix 1.

If you have a question in your mind which takes the form 'What would happen if we did such-and-such a thing?', the natural way to find the answer is to try it and see. This is the essence of an experiment; you take some action and you observe what effects it has.

It is possible to conduct an experiment without having any idea, in advance, of what is going to happen. Someone in distance teaching might say, 'Let's broadcast some radio programmes and see if anyone learns anything. Perhaps lots of people will listen attentively and learn a great deal, or perhaps no-one will learn anything. Let's just try it and see.' In general, however, you begin with a fairly clear idea in your mind about what will happen. You have a theory, or a proposal, and you want to test it. 'I think farmers will be more likely to adopt this new pesticide if they hear about it from an ordinary farmer who has already tried it rather than from an extension agent. Let's do an experiment to see if I'm right.' 'I think we should bring our correspondence students together for a one-day course at the beginning of their studies to give them some tips on how to study at home. Let's try it and see if it helps.' The technical term for what you think will happen is 'hypothesis'; the purpose of the experiment is to test your hypothesis, i.e. to see if you're right.

A good experiment needs to be planned carefully in advance; you should know exactly what action you are going to take, and you try to arrange things in such a way that you will be able to observe precisely what effects it has. Although the word 'experiment' tends to be associated in people's minds with laboratory work in physics or chemistry, experimenting is a method of research that can be applied just as well to practical questions in distance teaching.

Suppose that the student adviser of a correspondence college is worried about the proportion of students who drop out in the first year of their studies. To reduce the drop-out rate, he proposes that the college should send a standard letter of encouragement (a 'chivvy-letter') to any student who has not sent in any work for two months. The director likes this idea, but he is worried about the cost; he does not want to adopt this scheme as part of the college's routine without knowing whether it really makes any difference. So the student adviser decides to do an experiment.

The college is about to enrol its annual intake of 200 students. Each student is given a serial number as part of the regular administration of the courses. The student adviser gives instructions to the clerk responsible for communications with the students to send these chivvy-letters when appropriate to students who have even numbers but not to send any to students who have odd numbers. The college keeps regular records on the students' progress. At the end of the year, the student adviser compares the progress of the even-numbered students with that of the odd-numbered students. He finds that fewer of the even-numbered students have dropped out (23, as against 37 of the odd-numbered), and a statistician assures him that the difference is statistically significant (a term I explain in Chapter 11). He concludes that the chivvy-letters do make a difference.

This example shows that it is possible to do experiments in distance teaching, that the results of a good experiment can be clear cut and useful and that an experiment can be done without any great disruption in the organisation's regular work. But I don't want to give the impression that experimenting is easy. There are many traps for the unwary researcher, so you have to put a lot of thought into the design of an experiment, even a seemingly simple one.

'Internal validity': do the results really mean what they seem to mean?

Suppose that a course writer is in the early stages of writing a correspondence course, and he is wondering whether to include self-check exercises with each lesson. He decides to do an experiment to find out whether a lesson with self-check exercises is more effective than one without. He writes a lesson and he gets two different versions printed - one with exercises and the other without. Two of the teachers at the local secondary school are friends of his, and he persuades them to try out these lessons on their pupils. He prints enough copies for all the pupils and gives one version to one teacher and the other version to the other teacher. He asks them to get their pupils to work through the lessons in silence, in class. He also prepares a test paper which asks the pupils to give a written description of the main concepts of the lesson. The teachers are to give out these test papers after the pupils have finished reading the correspondence lessons. When the pupils have answered the test, the teachers are to collect the test papers, mark them and send the results back to him.

Everything proceeds smoothly. He delivers his lessons and test papers to the school and, the following week, the teachers send him their pupils' marks. He adds them up and finds that the class who had the version with the exercises did much better than the other class - averaging a score of 80% against the others' 45%. Very pleased at this striking result, he mentions it to the course editor at the correspondence college.

Now this editor does not favour self-check exercises and he is suspicious about this result, so he makes some enquiries. He discovers, first, that the pupils at this school are 'streamed'; that is to say, they are divided into separate classes according to their academic ability, the cleverest pupils in class A, the not-so-clever in class B, and so on. As it happened, class A got the version with the exercises and class B got the other. He also discovers that class A spent most of Tuesday morning doing their lesson and the test, whereas class B did it late on Friday afternoon, in a hurry. What's more, class A's teacher allowed them to keep their lessons while they did the test, whereas class B's teacher collected in the lessons before giving out the test. Finally, he discovers that class A's teacher is known to be generous in his marking, whereas class B's teacher is mean.

At first sight, the results of this experiment appeared to show conclusively that self-check exercises make a lesson more effective. But, on closer inspection, you can see that there are other explanations of the results. Given the above description of what actually happened, you would expect class A's marks to be higher, regardless of the self-check exercises. In fact, so many things went wrong with this experiment that you cannot tell whether the self-check exercises had any effect at all.

Experiments vary in the extent to which they are open to challenge. The experimenter maintains that the results show what he thinks they show, e.g. 'This shows that self-check exercises are effective.' His opponents argue that there are other explanations for the results, e.g. 'Class A were cleverer and they had more time and they could refer to the lesson while doing the test and the marking was more generous.' If the opponents can produce very plausible rival explanations, as in this example, the experimenter has a weak argument; his experiment is said to be weak in internal validity.

Consider again, by contrast, the example I described earlier about the chivvy-letters. The student adviser concluded that the results showed that the letters had some effect in preventing students from dropping out. If anyone wanted to dispute this, they would find it difficult to produce any plausible rival explanation of those results. The experiment, in fact, was carefully designed to isolate the effects of the letters and to exclude other factors that might interfere. An experiment which succeeds in excluding rival explanations is said to be strong in internal validity.

I have emphasised that the rival explanations have to be plausible. The reason why I have included the word 'plausible' is that, with some ingenuity, you can always produce some rival explanation for any experimental result. Even with the strong experimental design used in the chivvy-letters example, you can think of some other, possible explanations for the result. It is possible, you might say, that the clerk could see what result the student adviser wanted and was afraid he would get the blame if the result went the other way, so he sent out an extra letter of his own to all the even-numbered students, promising them a present if they stayed on the course. Or you might say that perhaps the students included a lot of married couples who enrolled together so that, as it happened, the odd-numbered students were predominantly husbands and the even-numbered were predominantly wives, and husbands tended to drop out more than wives, irrespective of any chivvy-letters. But these explanations are rather far-fetched. You could go on forever thinking up possible rival explanations and trying to rule them all out, it is only the plausible rival explanations that you need to worry about.

Some technical terms

An experimenter usually has a fairly clear idea about what is going to happen (e.g. 'I think the students who get the self-check exercises will learn better') and this is called his 'hypothesis'. If someone else offers an alternative explanation (e.g. 'Class A scored higher because they were cleverer, not because they had self-check exercises'), this is sometimes called a 'rival hypothesis'.

Many experiments are based on the idea of comparing two groups, as in both of my examples. There are those who get the 'special treatment' (the chivvy-letters, the lessons with self-check exercises) and those who don't (no chivvy-letters, lessons without exercises). Since education has taken these experimental designs from medical and agricultural research, the first sort of group - the ones who get it - is called the 'experimental group' or the 'treatment group', and the second sort of group - the ones who don't get it - is called the 'control group'. I don't much like these terms as they make teaching seem like a sort of vaccination, but they are widely used.

Another common feature of experiments in education is that students are given a test at the beginning of the experiment and then again at the end. If a course writer wanted to find out how much the students actually learned from a lesson, he might give them a test first, then have them work through the lesson, and then give them another test, so as to compare their scores on the 'before' test with their scores on the 'after' test. He might give them the same test twice, or he might use a different test for after. In either case, the test administered before is called the 'pre-test' and the one administered after is called the 'post-test'.

This sort of pre-test is quite different from the kind of pre-testing I mentioned in Chapter 1. The word 'pre-test', as I used it there, meant trying out a draft version of some educational material, such as a pamphlet, a poster or a correspondence lesson, to see if people understand it in the way they are meant to, before the thing is mass-produced; it is the material that is being pre-tested, i.e. tested before production. In the context of educational experiments, however, a pre-test is a test given to students before they receive the experimental treatment; it is the students who are being pre-tested, i.e. tested before the experiment.

The people on whom the experiment is carried out (the students, in my examples) are sometimes called the experimental subjects. This can be confusing since, in education, one is accustomed to using the word 'subject' in the sense of 'subject-matter', i.e. you might say that the subject of the lesson was geography or the theorem of Pythagoras; so, if you use the word in its technical sense, you should make this clear to the reader.

Comparing like with like

Both of the experiments I have described have hinged on comparing an experimental group with a control group. Not all educational experiments use this design, but a great many do and it is not hard to see why. You want to test the effect of some special 'treatment' (chivvy-letters, self-check exercises or whatever). You get two comparable groups of students similar in age, ability, educational level and so on - and you treat them in exactly the same way except that one group gets the special treatment and the other doesn't. If, at the end of the experiment, the 'treatment' group performs differently from the control group, this must be due to the special treatment. What other explanation could there be? In short, using a control group makes the experiment strong in internal validity.

However, this argument only holds to the extent that the two groups are in fact comparable. One of the weaknesses in the self-check exercises experiment was that the experimental group were cleverer than the control group, so their superior performance on the test didn't prove anything about the effectiveness of self-check exercises. But, you might argue, any two groups of people are going to differ in some respects. How can you arrange to have two groups who are completely comparable, or at least sufficiently comparable for an experiment?

The answer is to randomise, i.e. to assign subjects at random to the two groups. Suppose you had borrowed a class of forty schoolboys for an experiment and you wanted to divide them into two groups so that they would be evenly balanced in academic ability. One way to do it would be to take each boy in turn and decide which group to put him into by tossing a coin (or, being more sophisticated, by using a table of random numbers). This would give you two groups that were approximately equivalent, by which I mean that you wouldn't end up with all the bright boys in one group and all the dim ones in the other; you'd have an assortment of bright ones and dim ones in each group.

In Chapter 4, I explained how randomness is used to select a sample for a social survey, when you are using it to get representativeness. Now I am suggesting that you use randomness to assign people to groups; you are using it to get equivalence. The basic principle is the same; it's just that it is now being put to a different use.

A further reason for assigning subjects at random is that most of the statistical tests that have been devised for measuring the statistical significance of the results of such experiments are applicable only if the subjects have been assigned to the groups at random. (More about this in Chapter 11.)

If you were in a hurry, tossing a coin or using random numbers for each boy might take too long, so you might take a short cut. You could take a list of the boys in alphabetical order and assign them alternately to the two groups, or you might walk up and down the rows of desks assigning them alternately to the two groups. This would not be random assignment, strictly speaking, but it would probably be near enough. However, you do have to be careful about taking short cuts like this. A colleague of mine once divided a class of children into two groups for an experiment by drawing a line down the middle of the classroom. Only later did he discover that the children were seated in order of their academic ability - the brightest child at the front on the left and the slowest child at the back on the right. So, instead of getting two groups that were evenly balanced, he got exactly the opposite - the cleverer half of the class in one group and the slower half in the other.

If you could obtain some estimate of the boys' individual abilities in advance, such as their last exam results or their teacher's assessment, you could refine your random assignment by grouping the boys into pairs of approximately equal ability and then taking each pair in turn and tossing a coin to decide which member should go into the experimental group and which member into the control group. This procedure is known as 'matching'. This would give two groups that were better balanced than they probably would be from a simple random assignment, and it would eliminate the risk of getting a very bad balance - you couldn't possibly get all the cleverest boys in the same group.

Having got two comparable groups, you also have to make sure that they are treated in exactly the same way (except, of course, for the special treatment that the experiment is focussed on). Here again the self-check exercises experiment fell down; the experimental group had longer to do their lesson, they could refer to the lesson while taking the test, and their essays were marked more generously.

You can sometimes design an experiment so that both groups automatically get treated in the same way, as in the chivvy-letters example. Otherwise, there is no easy way to do it; you just have to be careful about all the details. If the experiment involves giving people a test, for example, you have to work out in advance exactly how the experimenter should administer the test, including what he should say to the students. This is especially important if different people are going to administer the test to different groups. To avoid the problem of bias in the marking of a test, it is better to design tests in which the answers are clearly right or wrong, by using multiple-choice questions, for example (see Chapter 13), or by designing a very strict and precise marking system.

If you borrow a class of schoolchildren for an experiment, practise the procedure in advance to get an idea of how long it will take, and allow yourself plenty of time to conduct the experiment in the classroom. You need to be calm and clear-headed when conducting an experiment, not hurried. Usually, you can ask the children to put their names on their test papers.

(In a social survey of the general public, it is better not to ask people for their names, as I explained in Chapter 5, but it is different with schoolchildren; they generally like to put their names on their test papers.) This can be useful if you are giving more than one test to the same class, or if you want to link up the test results to some other school records, such as marks in a previous school examination. You should also decide in advance whether you want some other items of information on the test papers, such as age and sex (if the sex is not always obvious from the name). Finally, both the schoolchildren and their teachers usually like to know the marks and, unless it would interfere with a further experiment you want to do with them, it's polite to let them know.

'External validity': will the results apply in real life?

Suppose that the course writer accepts that his experiment was too weak to support his conclusion that self-check exercises are effective, but he still thinks that self-check exercises might be helpful to the students. So he decides to repeat his experiment, except that, this time, he will avoid the mistakes of the previous one.

He keeps the same two versions of the lesson, i.e. one with self-check exercises and the other without. He rewrites the test so as to make it into multiple-choice form; there is no possibility of a marker being mean or generous when there is only one correct answer for each question. (The questions in the test are not the same as the questions in the self-check exercises.)

Again he gets some schoolteachers to help him; they all teach classes at the same level, but at different schools. (He does not use the same school as before.) This time, he asks only that they allow him to borrow their classes for two hours; he will conduct the experiment with each class himself. He obtains the list of pupils in each class before he goes to the school and, for each pupil, he tosses a coin to decide whether this pupil will get the version with the exercises or the version without. He plans carefully how he will actually conduct the experiment in the classroom; in fact, he writes down exactly what he will say to the pupils at each stage of the experiment and he practises reciting it.

He carries out the experiment with these classes, all of whom are at an appropriate level for this lesson. As a result of the coin-tossing, the pupils in each class are divided into roughly equal numbers doing the two versions. He takes care to conduct the experiment in exactly the same way with each class, starting at 10 a.m. on an ordinary school day each time.

He ends up with 140 test papers - 74 from pupils who had the version with the exercises and 66 from pupils who had the one without. He marks them himself and he also gets a colleague to mark them independently, as a check on the accuracy of his own marking. Finally, he adds up the marks and finds that the pupils who had the self-check exercises got a mean score (an average score) of 67%, while the others got 61%. He does a significance test (see Chapter 11) and finds that this difference is statistically significant at the 5% level. He concludes that self-check exercises do indeed make a difference, though not as much as he was hoping for.

Armed with this finding, he returns to the sceptical editor. Surely the editor cannot argue about this result? The course writer has taken care that the experimental and control groups were similar to each other (this was the purpose of the coin-tossing) and that they received the same treatment in every way except that one group had self-check exercises and the other did not. His experiment, in short, was strong in internal validity.

But the editor is not impressed. He agrees that the version with the exercises was more effective in this experiment, but he points out that the experiment was conducted with pupils at secondary schools, studying the lessons in ideal classroom conditions, in silence, and all doing what they were told to do. The real point at issue, he reminds the writer, is whether to include self-check exercises in a correspondence course. The people studying the correspondence course, in real life, will not be secondary school pupils. They will probably be working at home rather than in a classroom; they will probably not be working in silence, and they will probably not be doing exactly what the course tells them to do. Furthermore, says the editor, while conceding that self-check exercises made this particular lesson more effective, there is no guarantee that such exercises will have the same effect in other courses, or even in other lessons of this course.

Clearly, this editor is a hard person to convince, but his objection illustrates an important point. With any experiment, you can always say, 'It worked in this experiment, but will it work with other people or at other times, or in other places, or under different conditions?' In other words, 'How far can we generalise from this finding?' This aspect of experiments is known as external validity.

In practice, this basic question crops up in slightly different forms. One is when you are considering an experimental finding from elsewhere and you are wondering whether it will apply to your situation. The form of the question is, 'It worked there but will it work here?' Suppose you have found a research report showing that coloured diagrams make a lesson more effective than black-and-white diagrams, and you are wondering whether to adopt coloured diagrams in your next production, which is a simple guide for peasant farmers on the use of pesticides. In the original experiment, the test material was, let's say, a manual about cockpit controls and the students were trainee pilots of the United States Air Force. You cannot just assume that the finding will apply to your material and your audience; it might do, but it might not.

The question of external validity can arise even when you do your own experiment. In the example of the self-check exercises, the course writer wants an answer to a real-life problem ('Will self-check exercises help my correspondence students?'), but he conducts the experiment in a situation that differs from the real-life one, using secondary school pupils working in a classroom. The form of the question is, 'It worked in the experiment, but will it work in real life?'

Researchers often have no alternative but to accept this sort of discrepancy between the experiment and the real-life situation that they want to generalise to. Children who are attending school full-time are not, generally, the intended audience for a distance-teaching programme, but they are a convenient group to try things out on, simply because they are organised into classes and are accustomed to studying lessons and taking tests. Teachers are usually happy to let you borrow their classes for an hour or two, provided you don't do it too often. In a few days you can obtain experimental results from a hundred schoolchildren, whereas it would take weeks or months to conduct the same experiment with that many individual adults or correspondence students. The self-check exercises experiment illustrates this. Someone might argue that the course writer should write two different versions of his correspondence course and compare the progress of correspondence students working on the two versions. But such an experiment would require an enormous amount of work (producing two correspondence courses instead of one) and would take months or years to complete. The course writer wants an answer to his question before he starts writing the course, not two years after he has finished.

If the 'perfect experiment', as in this example, is impossible or impractical, you have to accept some loss of external validity, but you should still try to get as close as you can to the best experiment. Perhaps the course writer could find some people who were more like his future correspondence students, such as adults doing evening classes; it would be a bit better to conduct the experiment on them rather than on secondary school pupils. Perhaps he could ask them to take the lessons home and study them there, giving them the test when they next came to the evening class. Although this would have some disadvantages, it would be closer to studying by correspondence. (The main disadvantage is that he would lose control over what happened. Some of the students might work at home together and see each others' versions of the lesson, for instance.)

Even if the experiment resembles the real-life situation very closely, the question of external validity still arises if the people in the experiment know that they are taking part in an experiment. The problem is, simply, that people who know they are in an experiment behave differently from the way they would otherwise behave. This is often referred to as 'the Hawthorne effect'. It was given this name on account of some research conducted at the Hawthorne factory of the Western Electric Company in the United States. Some researchers were trying to see how far changes in working conditions (number of rest periods and so on) affected the workers' productivity. They discovered that these workers steadily increased their output throughout the period of the experiment, just because they knew they were taking part in an experiment. This effect was so strong that it completely obscured any effect that the changes in the working conditions might have had.

The Hawthorne effect is a particular problem if you are trying to assess the effect of some new educational material or technique by comparing it with the old one. The students who get the new material will apply themselves with more enthusiasm and will therefore learn better from it, just because they are in an experiment, regardless of the quality of the actual material. The question of external validity here takes the form, 'It worked when it was presented as part of an experiment, but will it work when it's in ordinary, everyday use?'

The way to avoid the Hawthorne effect is to arrange things so that the people do not know they are in an experiment. Often it is impossible (and perhaps unethical) to do an experiment on people without their knowing, but there are times when it is possible and not ethically objectionable. In the example of the chivvy-letters, the even-numbered students would not know that they were being treated any differently from ordinary students, and I don't think anyone could say that the college was treating the students unfairly by conducting this experiment on them.

Balancing internal and external validity

To recap, you can be confident, or doubtful, about whether an experiment really proves what the experimenter claims it proves; you are assessing its internal validity. And you can be confident, or doubtful, about whether something which has been shown to work in an experiment will also work in some other situation; you are assessing its external validity. An experiment can be strong in one and weak in the other. The chivvy-letters experiment was strong in both. The first self-check exercise experiment was weak in both. The second self-check exercise experiment was strong in internal validity but weak in external.

Unfortunately it often happens that the requirements of internal and external validity are in conflict; you can make your experiment stronger in one only at the cost of making it weaker in the other, and you have to strike a balance.

In academic research, experimenters are primarily interested in testing a hypothesis; they are not generally thinking of applying their findings in the real world. Consequently, they tend to pay great attention to internal validity and to ignore external validity. A university psychologist, whose special interest is the way people behave in small groups, might want to see whether people cooperate better in a group if the members are all of the same sex or if the sexes are mixed. So he conducts a carefully designed experiment on fifty university students who have volunteered to take part in psychology experiments. It is possible (in fact quite likely, I would guess) that these volunteers are not typical of people in general, or even of students in general, so that the findings from this experiment might not apply to other groups of people. So long as the researcher is not called upon to make any recommendation on the basis of his findings, this weakness of his experiment (i.e. in external validity) need not trouble him. And in fact a large number of experiments like this are conducted and reported in scientific journals. But suppose that the manager of a local factory asked the researcher for advice on the best way to organise his employees into small groups. The researcher would then have to face the possibility that the factory workers might behave differently from his student volunteers.

This lopsided character of experiments in academic research (internally strong, externally weak) makes them a bad model for a practical researcher to follow. The practical researcher should have a clear idea of the way in which his colleagues will want to generalise from his results, so he should design experiments which will allow the results to be generalised in that way. He will want more external validity in his experiments, even if it means losing some internal validity.

It is sometimes argued that internal validity is prior to external validity. In a sense, it is. If something has been demonstrated convincingly by an internally strong experiment, it is at least possible that the same thing will occur in other situations. Conversely, if the experiment is so internally weak that it does not prove what it purports to prove, the question of external validity does not even arise. In other words, if the findings are totally unreliable, you don't want to generalise from them at all. But there can still be a strong case, in practical research, for accepting some weakness in internal validity, if it gives more strength in external.

An example of this is provided by an educational institute which was attempting to evaluate the effectiveness of some number games in helping children to do arithmetic. The games were being used in a casual way in evening classes for out-of-school children. A social scientist who was asked to evaluate the effectiveness of these games designed an experiment. Some schoolchildren were recruited specially for the experiment. They were given a pre-test of their arithmetic skills, then divided randomly into experimental and control groups. One group played the number games intensively for an hour while the other group played word games. Then the children took the same arithmetic test again. The pre-test scores were subtracted from the post-test scores and the results indicated that the children who had played the number games had improved their arithmetic slightly more than the others.

This experiment used a classical experimental design, incorporating pre-tests and post-tests and assigning subjects randomly to the experimental and control groups. It was, therefore, strong in internal validity. You could say with confidence that playing the number games helped the experimental group to do better on the post-test; there is no other plausible explanation for the result.

But how much did the experiment tell the institute about the effectiveness of the number games as they were actually used in real life? Very little. The children who were recruited for the experiment were probably not typical of the sort of children who attended the evening classes. Playing the number games intensively for a whole hour was probably not how the games were played in the evening classes. The children certainly knew they were taking part in an experiment and therefore probably gave the number games more attention than ordinary children in an evening class would do ('the Hawthorne effect'). Finally, the institute presumably wanted to know whether the games had lasting effects on the children, but the experiment only showed that the children were better at arithmetic immediately after playing the games; there is no guarantee that the slight improvement in the children's arithmetic would have lasted very long after the experiment.

An experiment that gave more weight to external validity would have been more useful. Perhaps the researcher could have found some evening classes where the number games were not being used, or could have arranged that word games be substituted for the number games for a while in some of the classes. Then he could have given the pre-test to some of the children attending classes where they did have the games (the experimental group) and to some attending classes where they didn't (the control group). A few months later, he could give the post-test to the same children.

This experiment would be weaker in internal validity. If the researcher used some classes which did not have the number games, these classes might be different in other respects also - they might be more remote, or in poorer areas, or have less enthusiastic teachers - and these differences, rather than the number games, might explain any difference in the post-test results. Or perhaps, when he came to administer the post-test, he might find that many of the pre-tested children had stopped attending the classes. If this was more of a problem in one group than the other, this could produce a spurious difference in the results. But despite these weaknesses, this experiment would still have been more useful to the institute because of its superiority in external validity.

The aim of practical research is to give guidance. Experiments on practical questions do not have to be designed to yield conclusive results, though it is all to the good if they can be. It is more important that they should provide evidence which, if not conclusive, is at least relevant to real-life problems.

Real-life experiments, action research and pilot projects

Obviously, an experiment is strongest in external validity if it is conducted as part of a real-life operation. In the chivvy-letters experiment, the letters were sent out exactly as they would have been if they had formed part of the college's routine procedure. The student adviser could be confident that chivvy-letters in future years would have the same effect as in the experiment (though he could not be quite so confident if he had any reason to think that this year's students were different from usual).

Another example of a real-life experiment is a comparison we made at LDTC between slips of paper and stamped, addressed postcards. We had printed the first 10 000 copies of a cookery book and we were wondering what topics to choose for future booklets. We decided to include a slip of paper inside each copy of the cookery book, asking readers to write down what further topics they would be interested in and to send the slip of paper back to us. Someone suggested that stamped, addressed postcards would be better than slips of paper, but, of course, they would be more expensive. We decided to try both, as an experiment.

We arranged the booklets into groups of ten and put slips of paper into nine of them and a postcard in the tenth. Every batch of booklets was thus arranged so that every tenth buyer would get a postcard and the other nine would get slips of paper. This meant, virtually, that the postcards were distributed randomly among the buyers.

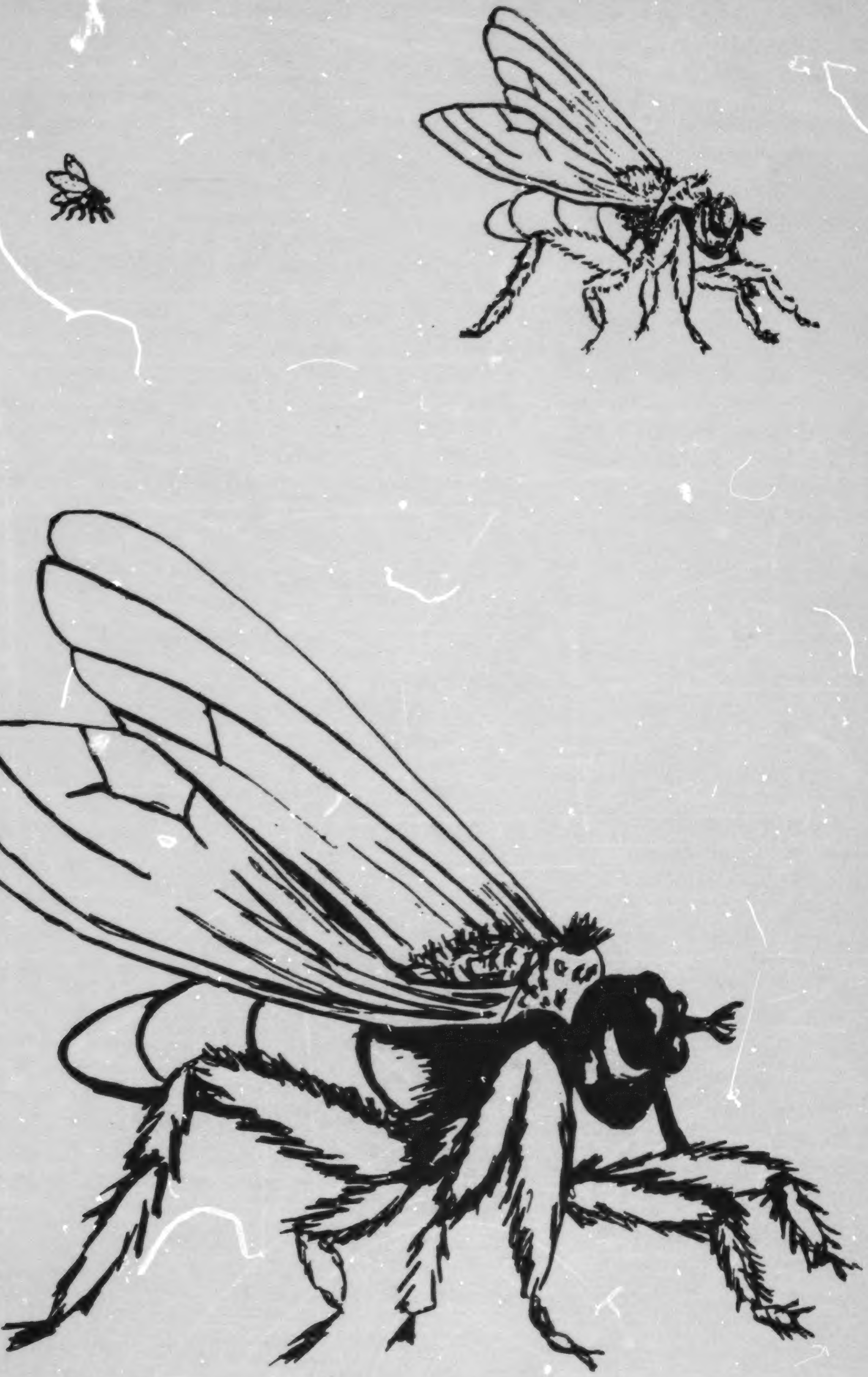
The booklets went out in November and we counted up the slips and postcards that we had received by the end of January - 362 slips of paper and 44 postcards. So, in a little over two months from the distribution of the booklets, 4.0% of the slips of paper had been returned (362 out of 9000) and 4.4% of the post cards (44 out of 1000) - a very small difference, not statistically significant. We concluded that the stamped, addressed postcards were no better than the slips of paper.

If it is impossible to incorporate an experiment into a real-life operation, you can sometimes increase external validity by incorporating it into a social survey. We conducted a survey of rural people in Lesotho, primarily to find out how many of them could read. Since it was costly simply to locate and interview a representative sample of rural adults, we thought we would get more for our money, so to speak, if we found out about some other things at the same time, so we built several experiments into the survey. The basic design was simple. Instead of having one set of test materials (pictures to look at, passages to read) for all the respondents, we prepared three slightly different packs of test materials. The interviewers used a different pack each day (pack A on the first day, pack B on the second, C on the third, A on the fourth and so on). The fieldwork went on for several weeks, so we ended up with our total sample of respondents divided into three similar groups.

An example will make this clearer. We had read that rural people, especially illiterate ones, do not recognise a magnified picture of something which is in reality quite small, such as a mosquito or a housefly. To test this, we prepared three different pictures of a housefly as in the illustration on the next page. The question was the same for all the respondents - 'What is this a picture of?' Those who happened to be interviewed with pack A got the life-size fly to look at; those with pack B got the 7 centimetre fly and those with pack C got the 17 centimetre fly. The results were striking. Of the 76 people who got the life-size fly, 61% recognised it correctly; of the 92 people who got the 7cm fly, 47% recognised it, and of the 77 people who got the 17cm fly, only 27% recognised it.

We could confidently assume that this result would apply to all rural people in Lesotho, since the subjects of the experiment, being our survey respondents, had been selected as a representative sample of that population. If we had conducted the experiment on a different group of people, such as schoolchildren in the capital city, the results would probably have been different and we could not have generalised confidently to the rural population.

Another type of real-life experiment is when you have an idea for a project and you test it by actually putting it into practice. This is sometimes called 'action research', meaning that it is an action project being undertaken as an experiment. LDTC's booklets scheme, in its first phase, was a piece of action research. We thought it was likely that, if practical booklets were made available at a reasonable price, people would buy them, read them and learn something useful from them (or, to be more precise, that enough people would buy them and that enough of these buyers would learn from them to make the project worthwhile). We had done research to find out how many people could read and what sort of reading matter was already available



Drawings of a fly

in the local language, but the only way to find out if our idea was right was to publish some booklets and to see what happened.

The research part of an action research project is much the same as a certain kind of evaluation, which I'll cover in Chapter 15. But, before leaving the topic of experiments, I should say a bit about pilot projects, since a pilot project is a sort of experiment.

The expression 'pilot project' is much misused. An agency which has grandiose ideas but limited resources will sometimes embark on a small scheme - a trimmed-down version of what it would really like to do - and will call it a pilot project. But it has no funds to follow this with a larger project, nor is it likely to get them. The pilot project finishes and that is the end of it. In fact, it never was a pilot project, just a small project.

A pilot project, properly speaking, is a trial run of what is intended to be a larger project. An agency which is intending to run a national campaign, for example, might try a small campaign first in just one part of the country. The idea is that they evaluate the pilot carefully; if it is successful, they then conduct the full campaign throughout the country, perhaps modifying the approach on the basis of the pilot experience. A pilot project is really a piece of research in itself, a small-scale experiment with an approach that is intended, later, to be reproduced on a much larger scale.

I emphasise the experimental nature of a pilot project, since there is a danger of the research side getting forgotten. It is easy to see how this happens. Naturally, people who are running a pilot project (the action people) want it to succeed, so they are inclined to put into the project any resources that are needed to make it succeed. Making modifications to the pilot project as it goes along is quite reasonable, up to a point; that is partly what the pilot is for. But it becomes unreasonable if they start putting resources into the pilot which they could not possibly put into the full-scale project that is intended to follow. This severely weakens the experimental value of the pilot project. As a piece of action, it might succeed; as a piece of research, it is flawed.

An example from LDTC will illustrate this. The rural education staff decided totry to encourage women's groups to use LDTC's educational materials. As an experiment, they contacted the leaders of eight women's groups and offered to help them teach their members how to crochet, using instructional materials produced by LDTC. This was to be a pilot project. If it proved successful, then LDTC would use this approach in the future with, perhaps, several hundred women's groups throughout the country.

The experiment proceeded smoothly, with the research department monitoring the groups' progress. Then, as the groups were about half-way through the course, the rural education staff suddenly decided that the groups needed some encouragement, and offered to give three balls of wool to every group member who completed the course. This was certainly not part of the experimental design. Furthermore, while LDTC could afford to give wool to the eighty women participating in the pilot project, it could not afford to give wool to the thousands of women who might eventually participate in a full-scale project.

The wool had the desired effect, at least from the point of view of the rural education department; the groups were flooded with new members wanting to join the crochet course. As a piece of action, the pilot was a success. But, as a piece of research, it had been sabotaged. The researcher had to try to disentangle the effects of the instructional materials, which could later be supplied on a national scale, from the effects of the wool, which could not. *

In this example, it is easy to see the extra resources that the action people were putting into the pilot, namely the balls of wool. People often make the same mistake in a less obvious way. Suppose that an organisation sets up ten study groups as a pilot project. With this small number, it is possible for the organiser to visit each of them quite often, thus giving each of them a lot of personal advice and encouragement. But he could not do this with a hundred groups. The extra resources that he is putting into the pilot, in this case, are his personal visits. It would be quite reasonable for the researcher to insist that the organiser visit the pilot groups much less often. The organiser might argue that the study groups might cease to meet if he did not visit them. But the researcher can reply that the purpose of the pilot is not to support ten study groups but to find out whether, given only the support they could expect in a full-scale project, the study groups continue to meet or not.

It is a question of balance. On the one hand, it is not reasonable or desirable to keep the action people strictly confined to a pre-arranged plan; they learn things in the course of the pilot and want to make adjustments as they go along. But if the effect of the adjustments is to turn the project into something that could not possibly be replicated on a larger scale, the pilot is no longer serving its function as a test run.

---

* The researcher in this (true) story was a colleague of mine. The rural education staff took the decision about the wool while he was in bed with tick-bite fever. Needless to say, he was furious when he got back on the job.

11 Statistical concepts

You need a few basic statistical concepts in order to analyse, or understand, the results of surveys and experiments. This chapter introduces these concepts. Most of the calculations and the background theory, however, are given in Appendices 1 and 2.

Readers who are not familiar with these concepts may find it best to take the chapter in two parts - the first three sections, which deal with results that can be expressed as percentages, and then the rest, which deal with results that are often expressed as averages.

Tables of percentages How to calculate percentages. Displaying results in percentage form. Base totals.

Confidence limits A survey result is an estimate of the true figure for the population. Confidence limits give you an idea of how wide of the mark your result might be.

Statistical significance There is always a possibility that your results are due merely to chance. A significance test tells you how serious this possibility is.

Category variables and measurement variables Certain procedures can be used with results in measurement form which cannot be used with results in category form.

Mean and median (averages) How to calculate them and what they tell you.

Variation Ways to describe the variation in a set of measurements.

The shape of a distribution Patterns in sets of measurements.

Statistical significance with measurement variables How the variation and shape of a distribution affect statistical significance.

The significance of statistical significance Results that are statistically significant are not necessarily significant in the ordinary sense of the word, and vice-versa.

Many people become nervous when they hear the word 'statistics'; they think that anything involving numbers is too abstract and complicated for them to understand. I suspect this is partly the result of poor teaching in mathematics but it is also because social scientists often use statistics that are unnecessarily sophisticated. At the higher levels, of course, statistics is an abstract and complicated subject, but you do not need an advanced degree in statistics in order to analyse the results of a survey or experiment. You can do an adequate job with just a few basic statistical concepts.

In this chapter I describe some statistical procedures that social scientists need to perform, concentrating on the concepts rather than the calculations. That is, I describe the reasons for performing the procedures and I explain what the results mean, but I don't show how to do the calculations, except for a few simple ones. If you want to actually carry out the procedures, refer to Appendix 1.

Tables of percentages

It is much easier to read a table of percentages than a table of raw figures. Take this cross-tabulation,* from a nutrition survey of mothers (I have invented the figures):

		Rural mothers	Urban mothers
Are you breast-feeding or bottle-feeding your baby, or both?	Bottle only	41	19
	Breast only	97	15
	Both bottle and breast	51	27
	Totals	189	61

Do urban mothers bottle-feed their babies more or less than rural mothers? It is not immediately clear from the table. But the answer is perfectly clear if you convert each column into percentages, like this:

		Rural mothers	Urban mothers
		%	%
Are you breast-feeding or bottle-feeding your baby, or both?	Bottle only	21	31
	Breast only	51	25
	Both breast and bottle	28	44
	Base totals	(189 = 100%)	(61 = 100%)

In order to convert a column of raw figures into percentages you add the figures together to give you the 'base total'. The base total of rural mothers, in this example, is 189. To calculate 41 as a percentage of 189, you do a bit of arithmetic:

$$\frac{41}{189} \times 100$$

This is very easy to do if you have a calculator. Remember that you divide the smaller number by the larger one, i.e. you calculate 41 divided by 189, not 189 divided by 41. If you can't remember how to do it, try working it out with some simple figures. For instance, 30 as a percentage of 60 should come to 50%, obviously. This is what you'll get if you calculate $(30 \div 60) \times 100$. If you get an answer of 2, or 200, you'll know you've done it wrong.

---

* The term 'cross-tabulation' is explained on page 77.

'Per cent' means 'per hundred'. When you calculate a column of percentages which ought to add up to 100%, you might have to round the percentages up or down so that they do in fact add up to 100, not 99 or 101.

Some people present percentages calculated to one or two decimal places; for example, they would present 19 out of 61 as 31.1% or even as 31.15%. This can be misleading for the reader. The way that you present a figure implies a certain level of accuracy. If you looked at your watch and said, 'The time is 32 minutes and 14 seconds past ten o'clock,' this would imply that you thought your watch was very accurate. It would be misleading to say this if in fact you knew that your watch was inaccurate and that all you were entitled to say was, 'It's about half past ten.' Similarly, if you have interviewed 61 urban mothers and found that 19 were bottle-feeding their babies, all you can really say is, 'About a third of urban mothers bottle-feed their babies.' Even to pin it down as 31% is slightly misleading, since it implies that your result is accurate to the nearest one per cent, which it isn't. But to say '31.1%' or '31.15%' is absurd.

Percentages can be particularly misleading when the base total is small. 'Seven respondents kept ducks, and two of the seven were women,' should not be reported as, '29% of the duck-owners were women.' In fact I would recommend that you do not present results as percentages if the base total is below 20.

There are various ways of setting out columns of percentages in a report. I suggest that, when the figures add up to 100%, you should put a % sign at the top of the column and the base total at the bottom, as in the table on the last page. Sometimes, the percentages do not add up to 100%, for example when the answers are multi-coded (see Chapter 7) or when you are condensing the results of several questions. In these cases, it is helpful to the reader of the report if you indicate that the figures do not add up to 100% by using a slightly different layout, such as this:

Who would you go to for advice on this problem? (250 = 100%) (Respondents could give more than one answer.)	Hospital doctor	73%
	Clinic nurse	54%
	Traditional healer	29%
	Mother-in-law	37%
	Other	21%

In general, if there might be any doubt about what a table of figures shows, you should add notes to the table to make it clear.

You do not always calculate percentages out of the total sample; often it is more appropriate to calculate percentages out of a particular subset of the total sample. Suppose you interviewed 220 married women aged between 20 and 70 about family planning, and you found that 53 of them were using a contraceptive method. It would be misleading to report that 24% of married women (i.e. 53 out of 220) were using a contraceptive method. It is only women in their childbearing years (up to age 45 or 50, say) who might need a contraceptive method. Say the sample contained 125 women aged 20 to 45: it would be more appropriate to report that 42% (53 out of 125) of married women aged 20 to 45 were using a contraceptive method. As another example, suppose you interviewed 120 people and found that 50 kept poultry and ten had vaccinated their birds. It would be misleading to say, 'Only 8% (10 out of 120) had vaccinated their birds,' when 70 people had no birds at all. You should say, '20% of poultry owners (10 out of 50) had vaccinated their birds.' When you present percentages in a report, always make it clear what group of people you have based the percentages on (i. e. who the people are who add up to 100%) and give the base total.

The following illustrations show how you might present tables of percentages in a survey report.

		%
Do you have any poultry?	Yes	52
	No	48
Base total (heads of households)		(126=100%)
		%
How many birds do you have?	One or two	29
	Three or four	27
	Five to ten	21
	Eleven to twenty	14
	Over twenty	9
Base total (poultry owners)		(66=100%)
		%
Have you asked anyone for advice about your birds in the last three months?	Yes	57
	No	43
Base total (poultry owners)		(64=100%)
Who did you ask?	Agricultural demonstrator	38%
The base total (=100%) are the 38 respondents who had asked for advice. Some gave more than one answer.	Official of Egg Cooperative	26%
	Poultry food salesman	18%
	Friends or neighbours	54%
	Other	12%
Note: Those few respondents (no more than three for any question) for whom we have inadequate information have been excluded from the base totals.		

		Lowlands %	Mountains %
Do you have any poultry?	Yes	60	37
	No	40	63
Base totals (heads of households)		(83=100%)	(43=100%)

Confidence limits

Suppose you interviewed a random sample of 120 heads of households and found that 66 of them (55%) owned poultry. If the sampling had been done properly, and if the checks on the data indicated that the results were fairly reliable, you would conclude that about 55% of all heads of households in the country owned poultry. The phrase 'about 55%' suggests some vagueness; how accurate is this figure?

Suppose that another organisation had carried out an identical survey at the same time. They too had interviewed a random sample of 120 heads of households - not the same people as you have interviewed, but another lot of people selected in the same way. You would not necessarily expect that they too would find exactly 66 poultry owners (55%). In fact you would be slightly surprised if they did find exactly 66. At the same time you would expect their result to be fairly close to yours. If they found that only 20% of their sample owned poultry, you would suspect that either their survey or your survey was wrong. In other words, different random samples, each of 120 people, would not give exactly the same answer to the question 'What proportion of people own poultry?' But the results should be fairly close.

This variation in result between one survey and another arises because each result is based on a sample of heads of households; one random sample will be slightly different from another random sample. Your survey result ('55% own poultry') is an estimate of the figure for the population, and it might be slightly wide of the mark; perhaps only 51% of heads of households actually own poultry, or perhaps it's 58%. This slight inaccuracy or vagueness in your result is called the 'sampling error'. The word 'error' here does not mean that anyone has made a mistake or done anything wrong; it just means that a survey result, being based on a sample, is only an estimate of the true figure for the population and it might be slightly wrong.

Note that this is different from the types of error I described in Chapter 8, such as errors that arise from interviewers not carrying out the instructions correctly or from some people refusing to be interviewed. All those other ways of getting a wrong result are sometimes grouped together in the expression 'non-sampling error'. Even if a survey was conducted perfectly, in the sense that all the interviewers did exactly as they were told and everyone cooperated and so on, the results might still be slightly wide of the mark because they would be based on a sample. Good training and supervision can help to reduce non-sampling errors, but sampling error is unavoidable.

However, though sampling error cannot be avoided, it can be calculated. That is, if someone asks, 'How accurate are the results?' you can give a precise answer, at least as far as the sampling error is concerned. In order for you to do this, the sample has to be a random one. As I mentioned in the chapter on sampling, this is where random samples are superior to other kinds of sample. With a random sample, you have a way of estimating how wide of the mark your result is likely to be; with a non-random sample, you don't.

An answer to the question, 'How accurate is this result?' is often given in the form of 'confidence limits'. If you said that the time was about half past ten and someone asked you how accurate your watch was, you might say, 'The time is certainly between 10.25 and 10.35,' or, if you were less confident about it, 'I guess it's between 10.15 and 10.45.' Similarly, if you were asked to say what you meant by the phrase 'about 55%', you might say, if you were very confident of the figure, 'The percentage of all households that own poultry is somewhere between 53% and 57%,' or, if you were less confident about it, you might say 'somewhere between 40% and 70%'. These upper and lower estimates are called 'confidence limits'.

To calculate confidence limits for a survey result, you first calculate something called the 'standard error'. I present the calculations in Appendix 1. I give the example there of a survey of 172 correspondence students selected at random from a large number of correspondence students. The survey has found that 81% of the sample are studying in the evening by candlelight. The standard error of this result comes out as 3%. What does this mean?

For reasons that I give in Appendix 2, there is a good chance that the true figure for the population falls within one standard error of the sample result, and a very good chance that it falls within two standard errors of the sample result. In this example, the standard error is 3%, so there is a good chance that the proportion of all the correspondence students studying by candlelight is within 3% of 81%, i.e. between 78% and 84%, and there is a very good chance that it is within 6% of 81%, i.e. between 75% and 87%.

To be more precise, by 'a good chance' I mean about two chances out of three; by 'a very good chance' I mean 95 chances out of 100. To put it another way, you can be 68% confident that the true figure is within one standard error (up or down) of your sample result, and 95% confident that it is within two standard errors. You could describe the figures of 75% and 87% in this example as your '95% confidence limits' for this result.

The thing that has most effect on the size of the standard error is the size of the sample. This is illustrated in the following set of figures, where I have calculated the standard error for a result of '70%' for samples of different sizes:

Result	Sample size	Standard error	95% confidence limits
70%	50	6.5%	57% - 83%
70%	70	5.5%	59% - 81%
70%	100	4.6%	61% - 79%
70%	140	3.9%	62% - 78%
70%	200	3.2%	64% - 76%
70%	300	2.6%	65% - 75%

As you can see from these figures, you can improve a small sample (of under 100, say) quite a lot by adding twenty or thirty cases. But when you have a large sample (say over 250), you have to add a hundred cases or more to make a worthwhile difference to the standard error.

Before finishing this section, I want to emphasise two points that I made earlier. The first is that the mathematical reasoning behind these calculations is based on the assumption that the sample is a random one. It would be nonsense to do these calculations for a sample that was not random. Secondly, the standard error for a random sample result gives you some idea of the amount of inaccuracy that is due to sampling error; it doesn't take any account of non-sampling error. If there was something wrong in the way people were selected or interviewed, or if a particular type of person had refused to co-operate in the survey, the sample result could be much further from the truth than the standard error suggests.

Statistical significance

Suppose there is a disease afflicting many people. Of those who catch the disease, about half die within a month of catching it, and half survive. A doctor invents a drug which, he thinks, will help to cure people of this disease. To test the effectiveness of the drug, he takes 160 people who have caught the disease and divides them randomly into two groups of 80. He gives the drug to one group but not the other. After a month, the results are as follows:

Table A

	Number of patients			Percentages	
	With drug	Without drug		With drug %	Without drug %
Survived	50	38		63	48
Died	30	42		37	52
Total	80	80	Base	80	80

Do these results prove that the drug works? Well, the drug obviously doesn't guarantee survival. 30 of those who took the drug still died. But a higher proportion of the drug-takers survived, so it looks as though the drug had some effect.

This conclusion, however, is not completely beyond doubt. On average, about half of those who get the disease survive, but out of any particular group of 80 sufferers, you would not necessarily expect that exactly 40 would survive. One group of 80 would differ from another group of 80. Out of one group perhaps 43 would survive, out of another group perhaps only 38, and so on.

In other words, you would expect some variation between one group and another just by chance. Occasionally you would get a group of 80 out of whom as many as 50 survived and, occasionally, a group out of whom only 30 survived. Returning to the doctor's experiment, it is possible that the group he gave the drug to just happened to be slightly exceptional; it is possible that 50 of them would have survived anyway, drug or no drug.

To make this point clearer, consider two other sets of results that the doctor might have got:

Table B

	Number of patients			Percentages	
	With drug	Without drug		With drug	Without drug
				%	%
Survived	72	38		90	48
Died	8	42		10	52
Total	80	80	Base	80	80

Table C

				%	%
Survived	39	38		49	48
Died	41	42		51	52
Total	80	80	Base	80	80

Table B would have demonstrated convincingly that the drug had an effect; surely such a large difference between the two groups could not be the result of random fluctuations. Table C, by contrast, wouldn't convince anyone; slightly more (in fact just one more) of the drug-takers survived, but the difference between the two groups is tiny and could easily have occurred by chance. The results he actually got are between the alternatives B and C. Do you find them convincing or not?

In considering this question, you are trying to assess the probability of getting certain results by pure chance. It is certainly possible that the doctor's results came about just by chance, but how probable is it? If the different results for the two groups could easily have arisen by chance, as in Table C, there is no reason to think that the drug had any effect. If it is most unlikely that the difference would have occurred just by chance, as in Table B, it strongly suggests that the drug did have an effect.

This idea of assessing the probability of getting certain results just by chance can be illustrated by a gambling game. Suppose someone invited you to play a game in which you took turns to throw three dice onto a table; if one player throws three sixes, the other has to give him some money. From time to time, someone will throw three sixes - roughly once in every 200 throws, in fact. But what would you think if your opponent did it three times in succession? You couldn't say that this was totally impossible; it could happen just by chance. But the probability is very small - less than one chance in ten million, in fact. You would be inclined to suspect that the dice were loaded or that he was cheating in some way. In other words, you would disregard the possibility of its being a chance result, because the probability of getting that result by pure chance is so small, and you would look for some other explanation.

To return to the doctor's experiment, the question is whether the probability of getting these results just by chance is so small that you can safely disregard it. Statisticians use the term 'statistical significance' when tackling these problems. If there is only a small probability that a certain result could have occurred just by chance, like the difference between the two groups in Table B, one says that the result is statistically highly significant. If the result could easily have occurred just by chance, like the difference between the two groups in Table C, one says that the result is not statistically significant.

The example of the doctor's drug has nothing to do with distance teaching, but the same question of statistical significance arises when one is analysing any results from a survey or an experiment. The following could be an example from distance teaching:

	Candidates who had taken a correspondence course	Candidates who had not taken a correspondence course
Passed the exam	50	38
Failed the exam	30	42
	80	80

Do these results show that the correspondence course helped students to pass the exam, or could the difference have occurred by chance? Or take the following result from a survey:

	Lowlands households	Mountain households
Had a radio	50	38
Did not have a radio	30	42
	80	80

Can you conclude that radios are more common in lowlands households, or is the apparent difference merely a chance result in your sample, not reflecting a real difference in the population?

Results like those I gave earlier in Tables B and C present no problems. The difference in Table B is obviously not a chance result; the difference in Table C could easily be a chance result. The question arises when the results are somewhere between these two extremes, like those that the doctor actually got. What one needs is some way of calculating exactly the probability of getting a certain result by pure chance. Statisticians have devoted much attention to this problem and have devised a large number of statistical tests, known as 'significance tests', which give a precise answer to this question. Given a particular set of results, a significance test tells you the probability of getting those results by pure chance. The test might tell you that the probability of getting those results just by chance is quite high - say one in three (33%) or one in four (25%) - or it might tell you that the possibility is quite small - say one in a hundred (1%) or one in a thousand (0.1%).

Suppose you did a significance test on the doctor's results and it told you that there was a one-in-five chance (20%) of those results occurring by pure chance. What then? You might argue that a 20% probability is quite small; the doctor's results were probably not due simply to chance, so they probably show that the drug was working. On the other hand, you might argue that 20% is too high for comfort. What if you were a health administrator trying to decide whether to recommend the drug for widespread use? You couldn't confidently adopt such a policy if there was a 20% risk that the experimental results were just a chance occurrence and that the drug was in fact completely ineffective.

The rule that has come to be widely adopted is that research results are convincing, or at least worth publishing, if the probability of their being due to pure chance is less than 5% - one in twenty. This is known as the '5% level of statistical significance'. For example, researchers in general would think that you should not publish results purporting to show that correspondence courses helped students to pass exams if there was more than a 5% probability that the results could be due simply to chance.

It is unfortunate, in some ways, that statisticians have chosen the word 'significance' to express the concept I have been explaining in this section. To say that the difference is statistically significant does not necessarily mean that the difference is large or important (i. e. 'significant' in the ordinary sense of the word). If you interviewed 1 000 men and 1 000 women about radio listening and you found that 532 women had listened in the previous week as against only 488 men, this difference would be statistically significant at the 5% level, but you would not say that the difference between 53% and 49% was an important difference. All it shows is that quite small differences can be statistically significant if the sample is large. Conversely, to say that a result is not statistically significant at the 5% level does not necessarily mean that it is of no interest. Say you obtained the following results from a survey of family planning:

	Table restricted to women aged 21-45 who had heard of family planning	
	Town women	Village women
Had visited an f. p. clinic	14	2
Had not visited an f. p. clinic	13	9
Totals	27	11

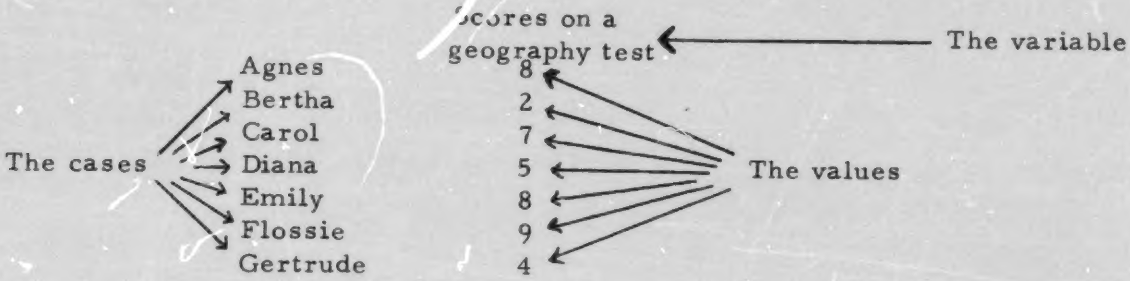
The difference here is not statistically significant at the 5% level, because the subsample is small, but the table certainly suggests that village women, even those who have heard of family planning, are less likely to visit a family planning clinic than town women. This might be worth including in the report, especially if you had other evidence to support it, though one should add a note to the table about the level of statistical significance.

Details of how to perform significance tests on the kind of results I've dealt with in this section are given in Appendix 1. Later in this chapter I return to the question of how to interpret significance levels in practical research.

Category variables and measurement variables

When analysing the results of a survey or experiment, the items you are talking about are called 'cases' and the characteristics are called 'variables'. If you have done a survey of poultry owners and found out how many birds each one owns, poultry owners are the cases and number of birds is the variable. The word 'variable' is used because the thing you are talking about varies from one case to another. One poultry owner might have 2 birds, another might have 10, and another might have 25; the number of birds varies from one case to another, so it is a variable. If you had details of the monthly earnings of 800 miners, and you were calculating the average monthly earnings, the miners would be your cases and monthly earnings would be the variable.

Mrs A keeps 2 birds, Mrs B keeps 8 and Mrs C keeps 27. The variable is the number of birds; the '2' '8' and '27' are called 'values' of the variable. Nicodemus got 5 out of 10 on the arithmetic test while Zebedee got 9 out of 10; the variable is the test score, and '5' and '9' are two values of that variable. If in doubt about these terms, refer to this diagram:



There are some characteristics of people that you can measure, and some that you can't. Age and height are examples of the first kind; age can be measured in years and height in centimetres. I will call these 'measurement variables'. The values of a measurement variable are measurements - 'This person is 25 years of age and 174 centimetres tall.' Sex and marital status are examples of the other kind of variable; you don't measure anything in order to describe someone as 'male' or 'married'. I will call these 'category variables'. The values of a category variable are categories - 'This person is a man and he's married,' or, 'This person falls into the categories "male" and "married".'

You can do things with measurement variables that you can't do with category variables. For instance, you can calculate the average. You could calculate the average age of twenty students but you couldn't calculate their average sex.

Do not be misled here by the fact that the categories of a category variable are sometimes given code numbers to make data-processing easier. Occupation is a category variable. Suppose you had details of the occupations of twenty correspondence students; you divided them into five categories and gave each category a code number, as follows:

Teacher	Code 1
Policeman	2
Civil servant	3
Farmer	4
Other	5

Each student's occupation is represented by one of these numbers, so, if you wanted, you could add them up and divide by 20 (the number of students); you might find that the average was 2.87. But this figure would be meaningless; it would be nonsense to say, 'The average occupation was 2.87.' When code numbers are used for categories they are used merely as labels, rather like the numbers on football shirts. If you perform arithmetical operations on them - adding, dividing and so on - the results make no sense.

Measurement variables can be converted into category form, but not vice versa. You might decide that you are not interested in the exact age of each student, only in their age-group, so you might recode their ages as follows:

20 or under	Code 1
21 to 30	2
31 or over	3

Whereas age, measured in years, is a measurement variable, this grouped version of it is a category variable.

Examination boards do this when they convert candidates' exact marks ('87', '56', '29') into grades (A, B, C, etc.) or into a simple 'Pass' or 'Fail'. Similarly, in a survey of farmers, you might record the exact number of cattle that each one owns (a measurement variable), but then categorise these in the analysis as follows:

1 - 5 animals - small holding
6 - 20 " - medium holding
21 - 60 " - large holding
over 60 " - very large holding

In converting a measurement variable into category form, you lose some information. You are throwing away the precise measurement of each farmer's holding ('16 cattle', '23 cattle') and replacing it with a label ('medium holding', 'large holding'). Farmers with 62, 87 and 110 cattle are no longer differentiated; they are all in the same category - 'very large holding'. This is why you can convert measurement variables into category form but not vice versa. If you collect precise measurements, you can group them into broad categories later, if you want to, but if you collect the information in categorised form, you can't convert this later into precise measurements.

In analysing survey results, you are mostly dealing with category variables, either because they are naturally in category form (sex, marital status, Yes/No answers) or because they are measurement variables that have been categorised (age-group, size of cattle holding). Occasionally you collect precise measurements and use them - to calculate the mean cattle holding, for example - but most survey analysis consists of sorting people into categories and counting them. Certain statistical procedures are appropriate for category variables but other procedures are required for measurement variables. In analysing the results of experiments, you are more often dealing with measurement variables, especially test scores.

Mean and median (averages)

Suppose you have found out, from eleven poultry owners, how many birds each one possesses, and the results are as follows:

2, 2, 2, 4, 4, 5, 9, 9, 12, 16, 23.

What is the average number of birds? You just add up the total number of birds and divide by eleven:

88 ÷ 11 = 8.

Statisticians use the term 'mean' to describe this kind of average. They would say that the mean of these eleven numbers is 8, or that the mean size of these eleven poultry holdings is 8 birds.

The reason for using the word 'mean' rather than 'average' is that statisticians use other kinds of average apart from the mean. One of these other averages is called the 'median'. To find the median, you put all the numbers in a line, from smallest to largest; the median is the number in the middle of the line. In the example of the poultry holdings, there are eleven numbers in the line, so the median is the sixth from the left (it's also the sixth from the right, i. e. it's in the middle). This number is 5. So the median size of these eleven poultry holdings is 5 birds.

If you had an even number of poultry holdings, you would take the two in the middle; the median would be halfway between them. For example, say you had these eight poultry holdings:

2, 4, 4, 5, 9, 9, 12, 16.

The median would be halfway between the fourth number and the fifth i. e. halfway between 5 and 9, so the median holding would be 7 birds.

The mean is used more often than the median. When people just say 'the average', it is usually the mean that they are talking about. However, the mean can sometimes be misleading. Suppose that the eleven poultry holdings were like this:

2, 2, 2, 4, 4, 5, 9, 9, 12, 16, 595.

Our example now contains ten poultry owners with fairly small holdings, and one with an exceptionally large holding. The mean of these eleven holdings is 60 birds.

Now, if you hadn't seen the actual figures and you were told that the average holding was 60 birds, you might get the idea that a holding of 60 birds was typical; you might think that if you called on a typical poultry owner, he would have about 60 birds. But you would be wrong. Actually no-one has a holding of about 60 birds. The typical poultry owner, to judge from these eleven, has just a few birds. In this example, the median (5) would give a better idea of the typical poultry holding.

Averages are very useful statistics, and people use them all the time - the average family size, the average household income and so on. But you must bear in mind that, if there is much variation in the figures, the average can give a false impression.

Variation

Suppose you gave the same test to two groups of students, nine students in each group, and they got the following scores (each one out of 100):

Group A	Group B
62	95
60	86
56	67
54	53
48	48
45	40
43	32
42	23
40	6

Did one group do better than the other? Well, some students in group B got very high scores, but some also got very low scores, so you can't say that, overall, group B did better. In fact, the mean score for group B is the same as for group A (50), and the median scores are also the same (48). How would you describe what it is that is different between these two sets of scores?

The scores of group A are clustered around the mean, whereas those of group B are more spread out. You could say there was more variation in the scores of group B than in those of group A. The mean and the median give you some idea of the mid-point in a set of scores, but they don't give any indication of the amount of variation. For this, you need some other figure.

One way to show the amount of variation in a set of scores is to calculate what is called the 'mean deviation'. You calculate how far away from the mean each of the scores is and then you work out the mean of these amounts. The distance between each score and the mean is called a 'deviation' so the average of these distances is the 'mean deviation'. For example, the mean score of group A is 50. The first score is 62, so the first deviation is 12; 62 is 12 points away from 50.

Here are the calculations for the mean deviation of group A:

Score	Deviation of the score from the mean
62	12
60	10
56	6
54	4
48	2
45	5
43	7
42	8
40	10
Total	64

Mean deviation = 64 ÷ 9 = 7.11 (to two decimal places)

If you did similar calculations for group B you'd find that the mean deviation was 22.44, which reflects the fact that the scores of group B are more spread out.

The mean deviation is a measure of variation that is easy to grasp, but social scientists hardly ever use it. This is because, for more advanced statistics, there is another measure of variation that is more useful. This other measure is called the 'standard deviation', often symbolised by the letter 's'. Appendix 1 describes how to calculate the standard deviation and Appendix 2 explains why statisticians make so much use of it.

To give you some idea of how the standard deviation reflects the variation in a set of figures, here are several possible sets of scores that nine students might have got, with the standard deviation calculated for each one (the mean is 50 for each set):

Group	C	D	E	F	G	H	I
	50	52	54	80	90	90	100
	50	52	54	54	45	80	100
	50	51	52	52	45	70	100
	50	50	50	50	45	60	100
	50	50	50	50	45	50	50
	50	50	50	50	45	40	0
	50	49	48	48	45	30	0
	50	49	48	48	45	20	0
	50	47	44	20	45	10	0
s =	0	1.49	2.98	14.25	14.14	25.82	47.14

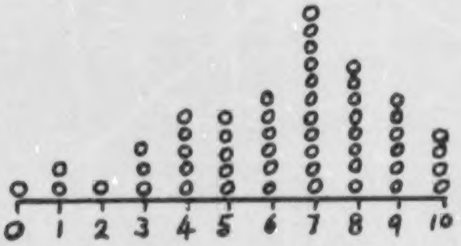
In group C, there is no variation at all, so the standard deviation is zero. Group D shows that a tightly clustered set of figures produces a small standard deviation. The standard deviation for group E is precisely double the one for group D, and this reflects the fact that each individual deviation has been doubled; 54, for instance - the first mark in group E - represents a deviation of 4 marks from the mean (50), as against the corresponding score of 52 in group D, which is only 2 marks from the mean. Groups F and G show how just one or two extreme figures in an otherwise tightly clustered set increase the standard deviation markedly. Groups H and I show that if a lot of the scores are a long way from the mean, the standard deviation is large.

The shape of a distribution

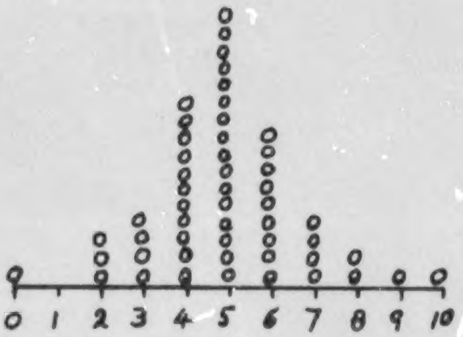
Suppose you gave a test to a class of 52 schoolchildren and they got the following scores, each one out of ten:

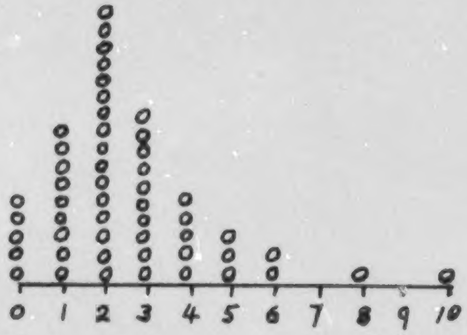
Anne	5
Arthur	1
Brenda	7
Brian	0
Clare	5
and so on, down to	
Yves	6
Zoe	1
Zebedee	10

Suppose that you drew a straight line across the school playground and wrote the numbers from 0 to 10 along the line and then positioned the children along this line, getting them to stand next to the number of marks they got in the test. Brian, for example, who got no marks, would stand next to the 0; Arthur and Zoe would stand next to the 1, and so on. Someone who happened to be passing over the playground in a helicopter, looking down, would see something like this:



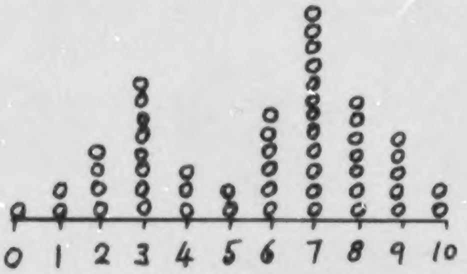
This diagram shows how many children got each score. To put it another way, it shows how the scores were distributed among the children, and for this reason, a set of measurements like this set of scores is called a 'distribution'. The diagram shows how the variable 'test score' was distributed among this class of children. As this diagram of the distribution makes clear, a distribution has a pattern or shape. Consider these two sets of scores that the children might have got:





The distribution on the left is symmetrical. The score occurring most frequently in the list was 5. This is called the 'mode' and it shows up as a peak in the diagram. The other scores are distributed fairly evenly on each side of it. The distribution on the right is not symmetrical. The scores are bunched to the left - the mode score is 2 - and the distribution has a long tail out to the right. A distribution like this is called a 'skew' distribution. The first diagram I gave also showed a skew distribution, the scores in that example being bunched to the right.

All three of these distributions have something in common - they are all 'unimodal', i.e. each diagram has only one peak. Occasionally, though not often, a distribution has more than one peak. The following is an example of a 'bimodal' (two-peaked) distribution:



Statistical significance with measurement variables

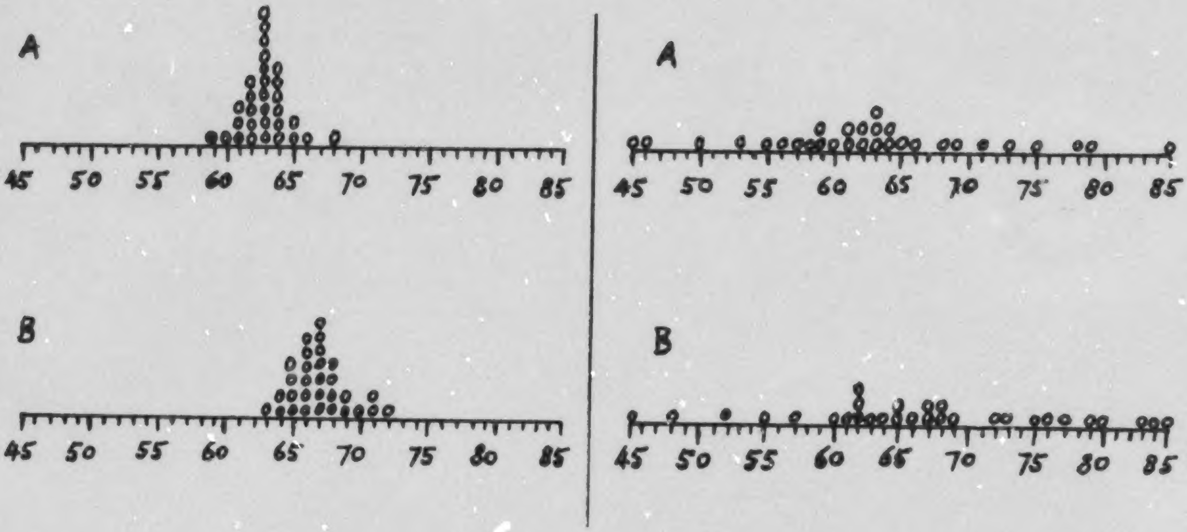
Earlier in this chapter I explained the concept of statistical significance with the example of an experiment whose results were expressed in terms of a category variable; at the end of the experiment the experimental subjects (the patients) were classified into just two categories - 'Died' or 'Survived'. The same question of statistical significance arises when the results are in the form of measurements.

Suppose a researcher tests two versions of a draft correspondence lesson, to see which version is the more effective. Let's assume that he avoids the mistakes I described in the last chapter, so that his experiment achieves a high degree of internal validity. The results of the experiment are in the form of scores on a test, i.e. measurements. The thirty students who take version A get a mean score of 63 out of 100, while the thirty who take version B get a mean of 67. Can he conclude that version B was more effective?

The researcher's problem is very like the one that faced the doctor in the earlier example. One group of thirty students is not exactly the same as another group of thirty, even though they have been assigned to their groups at random. Even if both groups had had the same version of the lesson, they would not necessarily have got exactly the same average score; one group would probably have done slightly better than the other just by chance. What is the probability of getting this result (67 against 63) just by chance? If a difference of this size is quite likely to occur by chance - with a probability of, say, four chances out of ten - the researcher cannot safely conclude that version B was really better. If, on the other hand, a difference like this is unlikely to occur by chance - say only one chance out of 100 - the researcher can say with confidence that version B really was more effective. The researcher needs a significance test to see whether the difference between the two mean scores is significant or not at, say, the 5% level.

What I've just said implies that the size of the difference between the means is the only thing that matters. Actually it's more complicated than that. The size of the difference between the mean scores is obviously one of the important things; if one group had scored 95 and the other group 36, this clearly could not have occurred just by chance (more precisely, the probability would have been extremely small). But the variation within the two sets of scores and the shape of the two distributions also come into it.

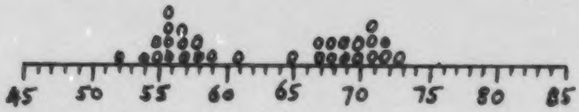
This diagram shows how the amount of variation in the scores affects the statistical significance of the difference between the two means. In both sets of results, group A's mean is 63 and group B's is 67:



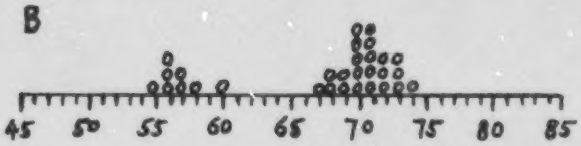
In the left-hand set of results, there is very little variation within the scores of each group - the standard deviation for group A is 1.73, for group B 2.11. As a result, the scores of the two groups are clearly different; most of the students in group B did better than the students in group A. In the results on the right, by contrast, the difference between the two groups is not nearly so clear. There is a lot of variation within each group - the standard deviation is 9.05 for group A, 10.00 for group B - so the two sets of scores overlap. Group A had slightly more low scores and group B had slightly more high scores, but the two sets do not look very different. Even without doing a significance test you would guess that the difference between groups A and B on the right could be due to chance, whereas the difference between the groups on the left is clearly not due to chance.

The shape of the distribution affects things in a different way. If the shape is out of the ordinary, you should probably look in more detail at the results rather than simply test the difference between the means. By an 'ordinary' distribution, I mean one in which the values are somewhat bunched together around the mean, with some tailing off to each side. All four distributions in the above example count as ordinary distributions. The following examples are out of the ordinary (as before, group A's mean is 63 and group B's 67):

A

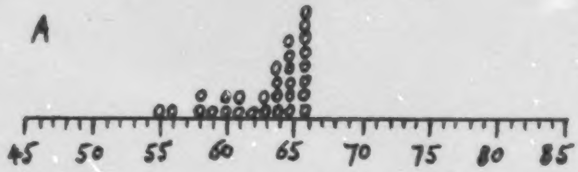


B

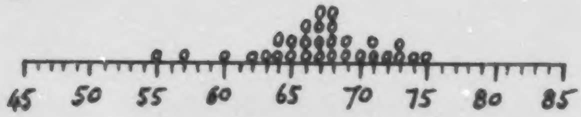


Something odd is going on here. Perhaps the sixty students were composed of two distinct types; for example, perhaps some had studied the topic of this lesson before and some hadn't. Or perhaps some of them had misunderstood some part of the lesson or the test. Before doing any significance test you ought to investigate the reasons for these bimodal distributions. It would be over-simple to say merely that group B had done better than group A. The following shows another way in which a distribution can be out of the ordinary:

A



B



The main thing here is not just that group B did better but that group A's distribution is skew whereas group B's isn't. Why did none of the students in group A, even the cleverer ones, score more than 66? Did their version of the lesson omit some information that was essential for answering some of the questions? Or did it actually mislead the students into giving the wrong answers? You could draw wrong conclusions by just proceeding with a significance test without finding out the reasons for group A's odd distribution.

Appendix 1 presents four significance tests for use with measurement variables and explains how to choose the one you want.

The significance of statistical significance

People who do experiments with an eye to publishing a report in a learned journal face a simple problem: are the results worth publishing or not? To make things easy and fair, the academic world has adopted a simple rule: results are worth publishing if they are statistically significant at the 5% level. Because of this convention, the 5% level has become endowed with an almost mystical authority which it does not deserve.

In accepting that results mean what they seem to mean, you are taking a risk; perhaps they were just chance results, not to be relied on. The significance level you adopt is the size of the risk you are prepared to take. In everyday life, the level of risk you are prepared to accept depends very much on the situation. I might go to a football match in the hope of getting in even though I knew there was a 5% chance of the stadium being full; I would be prepared to accept a 5% risk of disappointment. By contrast, if I went to the doctor's to get an injection for travelling overseas and he told me there was a 5% risk of my getting serious and possibly fatal side-effects from the injection, I'd refuse the injection and cancel the trip; in that situation, a 5% risk would be much too high.

If a researcher in a distance-teaching organisation does a quick experiment for a course writer, comparing two styles of lesson, the question that he and the writer face is not whether to submit the results to a journal for publication, but whether to be guided by the results in adopting a style for the course. Suppose the results suggest that style B is the more effective, but, because only a few students took part in the experiment, the difference is significant only at the 20% level. Should the researcher recommend style B to the writer, with reservations about the significance level, or should he discard the results because they fail to reach the 5% level? Ideally, perhaps, he should repeat the experiment, perhaps with a larger sample, but, failing that, the more reasonable course is, surely, to recommend style B. Admittedly, it has not been proved beyond all doubt that style B is superior, but the results provide more support for style B than for style A.

In practical research, the 5% level need not be given the reverence it receives in academic work. Results that attain only the 10% or even 20% level might still be useful.

Conversely, results that do attain the 5% level of significance are not necessarily reliable. If an experiment contains any of the faults I described in the last chapter, the level of statistical significance is irrelevant. Suppose group A get an average score of 40 and group B of 80. A significance test might well show that the difference is highly significant statistically, i.e. almost certainly not due to chance. But it doesn't necessarily prove that version B was more effective; perhaps the results only reflect the fact that group B were cheating. A significance test does not bestow a certificate of approval, so to speak, on every aspect of an experiment or survey.

And even reliable results that attain the 5% level are not necessarily valuable. I pointed out earlier that, if the sample is large enough, even small differences become statistically significant. Or suppose that some students did better on a test after working through a lesson than they did before, and that the pre-test/post-test difference is significant at the 5% level. All this shows, at best, is that some of them learned something from the lesson, and this is hardly an exciting finding; it's the least you'd expect.

In short, you should pay some attention to statistical significance, but don't let it dominate your thinking. Results that attain a high level of statistical significance can still be misleading or trivial; results that fail to attain a high level can still be useful.

12 Assessing instructional materials

This chapter describes a few ways of assessing the quality of instructional materials before you go to the trouble of pre-testing them on students. It is mainly about printed matter, especially materials in English, but the first two sections could apply also to radio programmes (in script or on cassettes).

Clarifying objectives *It is a great help in producing, assessing and testing materials if the producers make clear their objectives. Specifying objectives in behavioural form can be a good way of doing this.*

Collecting expert judgements *Circulating draft materials to experts for comment may indicate ways of improving the material.*

Measuring readability of English text *I describe how to calculate a Reading Ease score for a piece of English text. This gives a rough idea of how much difficulty readers would have with it.*

Distance teaching relies heavily on instructional materials - correspondence course texts, booklets, leaflets, radio programmes or whatever. Sometimes a student can turn to someone for extra help if he can't understand the materials - to a correspondence tutor, a radio group leader, or a fellow student, for example - but generally he is expected to gain most of his learning from the materials themselves. If the materials are poor, the students won't learn much. This chapter and the next - they can be taken together - are about using research to improve distance-teaching materials.

The best way to see if some materials are good enough is to try them out on some students. This is called 'pre-testing' and I'll describe it in the next chapter. But pre-testing requires some effort - producing several copies of the trial materials, contacting a sample of students, devising a test and so on - so you want to be sure that the trial materials are at least sufficiently good to be worth the trouble of pre-testing. This chapter is about ways of assessing instructional materials in advance of pre-testing, so that the materials can be improved even before they are tried out on students.

Clarifying objectives

Suppose that a course writer, writing a maths course, submits a draft lesson on logarithms to the editor. The editor reads it and then puts some criticisms to the course writer. 'Am I right in thinking that one uses logarithms in order to convert complicated calculations into simpler ones? In other words, they are an aid to calculation? I think that's an important point, but you haven't mentioned it in the lesson.' 'You say, near the end, that you can have logarithms with a base other than 10, but you don't explain what the word "base" means in this context.' 'The problems that you give the students as a practice exercise at the end of the lesson are of a different kind from the ones you've illustrated in the lesson.' Sooner or later, the editor will put the crucial question, 'Exactly what are you trying to get across to the students in this lesson?'

If a teacher puts together a lesson without having a clear idea of what he's trying to teach (and this applies to classroom teaching as well as to distance teaching), the lesson is likely to be inconsistent and incomplete. As a result, when the students come to take what is supposed to be a test on the lesson, they are not equipped to cope with it.

In order to assess whether some instructional material is likely to do its job adequately, you need to have a clear idea of the job that it's supposed to do. One way to get the producer of the material to be clearer about this is to ask 'How will you know if the students have learnt what's in the lesson? How will they show it? What do you want them to be able to do?' In the above example, it is not enough just to know that the lesson is supposed to teach the students about logarithms. At the end of the lesson, are they supposed to be able to solve problems by using logarithms? What sort of problems?

The aims of a lesson, described in terms of what the students should be able to do at the end of the lesson, are known as 'behavioural objectives'. In asking a teacher to describe his aims in this way, you are asking him to specify the behavioural objectives of the lesson.

The word 'behavioural' is important. If you asked the course writer to define the objectives of his lesson, he might say, 'I want the students to understand logarithms.' But you can't tell if a student understands logarithms just by looking at him. You have to ask him to do something. You'd have to give the student some problems, provide him with a set of log tables and ask him to solve the problems by using the tables. It is perfectly reasonable for the teacher to say, 'I want the students to understand logarithms. That is a respectable objective. But it does not specify what the students will be able to do, so it is not a behavioural objective.

When trying to specify objectives in behavioural terms, avoid words like 'know', 'understand', 'appreciate', 'grasp the significance of', since they do not describe precisely what the students will be able to do. Rather, you should favour words that describe actions, such as ' 'write', 'tell', 'identify', 'name', 'solve', 'construct', 'give a list of', 'describe'.

I have used an example from maths teaching, but the same principles can apply also to non-formal education. The writer of a pamphlet on pesticides might specify his objectives as follows, 'After reading the pamphlet, a farmer should be able to name at least two pesticides, to say what pests they are used for, and to describe at least three safety precautions that one should take when using pesticides.' The scriptwriter of a family planning radio programme might say, 'A listener should be able to suggest two ways in which family planning can be beneficial to a young couple.' Both of these are behavioural objectives, since they describe what the readers or listeners should be able to do ('name', 'say', 'describe', 'suggest'). The objectives would not have counted as behavioural objectives if the writers had expressed them as follows: 'I want farmers to appreciate the value, but also the dangers of pesticides,' or 'I want people to see the benefits of family planning.'

So far in this section, the only characters I have mentioned have been writers and editors. Where does the researcher come in? He comes in at the pre-testing stage. A draft of the material has been produced to be tried out on some students and the researcher has to devise a test to see if the students have learnt what they were supposed to learn from it. The question he has to face is, 'What would be an appropriate test?' and this leads immediately to the question 'What are the students supposed to be able to do at the end of the lesson?' So, if the writer has not already thought of the purpose of his material in terms of behavioural objectives, he has to do so at the pre-testing stage, in order that the researcher can devise an appropriate test.

In fact, it is better if the writer and researcher can get together well before this stage, perhaps even before the writer has begun writing. At the researcher's request, the writer tries to specify behavioural objectives and this helps him to write better material. The researcher drafts a test in the light of these specified objectives and shows it to the writer. The writer perhaps feels that parts of the test are inappropriate and tries to be even more precise about his objectives. The clarity that ought to result from this co-operation leads to both a better lesson and a better test.

To return to the example of the lesson on logarithms. The writer says he wants the students to understand logarithms. The researcher asks, 'Do you want the students to be able to solve problems using logarithms?' The writer says 'Yes' and immediately realises that the lesson should contain an explanation of how to get the right logarithm from the log tables, as well as how to use the logarithm once you've got it. The researcher drafts a few test problems, including '26.957 ÷ 196.2979'. 'Oh no,' says the writer. 'I'm only covering numbers less than 10, and with only two decimal places.' And so, together, they get a progressively more precise idea of what the lesson is trying to teach and they devise an appropriate test for it.*

When trying to make behavioural objectives more precise, it sometimes helps to specify the conditions under which the students will take the test. Are there any special aids that will be provided for them, or, on the contrary, is there anything they will be forbidden to use? In a maths test, for example, will they be allowed to use a calculator, or a slide-rule? In a language test, will they be allowed to use a dictionary? If it's a test for car mechanics, will they be given an engine in good order which they have to dismantle and reassemble, or will they be given a faulty engine which they have to repair?

Finally it is a good idea to consider, in advance, what will count as an acceptable level of performance. The maths teacher wants his students to be able to solve problems of a certain type using logarithms, but presumably

---

* In this example I have regarded the writer and the researcher as two different people, but a writer may want to pre-test his own material, in which case he will ask himself these questions. A good writer will have formed a fairly clear idea of his objectives, perhaps with help from an editor, even before beginning to write the material.

(unless he's very new to teaching) he doesn't expect all his students to solve every problem correctly. Will he be satisfied if they get three-quarters of the problems right? This question may be hard to answer as it focusses attention on the weaker students. The better students may well get all the answers right, but what if a few of the weaker ones do badly? How much is the writer going to worry about them? Similarly, the writer of the pesticides pamphlet cannot expect every farmer to commit to memory the entire contents of the pamphlet. What if it turns out that most of the readers remember the names of the pesticides but forget the safety precautions? Will the writer be happy with that or will he decide, in that case, that the pamphlet needs rewriting?

By saying in advance what will count as a minimum acceptable level of performance on the part of students who have studied his material, a materials producer is, in effect, committing himself to making changes to his material if it turns out that the students do not in fact reach the level he has set. 'I want the students to get three-quarters of the problems right. If they don't, I'll have to rewrite the lesson.' This makes it more likely that he will indeed make improvements to his material if the pre-test results suggest that he should. If he hasn't made such a specific commitment in advance, it is easy for him just to ignore the results.

Before leaving behavioural objectives, I should discuss briefly an objection which is often raised against them. Educators often resent being asked to specify their objectives in behavioural terms; they protest that they are not trying to influence the behaviour of their students, or at least not in a simple, direct, observable way. This emphasis on behaviour, they say, might be appropriate for certain subjects; if you're trying to teach someone to drive a car, for example, then it's appropriate to test his driving at the end. But it's clearly not appropriate for some other subjects. The main aim of a course in English literature, for example, might be to heighten the students' appreciation of poetry. That's quite different from teaching them a physical skill like driving.

It must be admitted that the word 'behavioural' is a bit misleading. Perhaps 'observable' objectives would be better, or 'testable' or 'measurable'. No one would suggest that a poetry appreciation course consisted of training the students to perform certain actions. But surely there must be some way of seeing whether a group of students, who didn't appreciate poetry much at the beginning of the course, appreciate it more at the end. How about giving them two poems and asking them to note points of similarity and points of contrast? If the course has had an effect, they ought to do this better at the end, noting more points or noting less obvious points. Perhaps the course writer would reject this suggestion, but at least it would get him thinking about how to assess the effectiveness of the course.

Collecting expert judgements

When a writer has just finished a draft, it is difficult for him to assess how good it is. He is too close to it, so to speak, and it takes time before he can look at it through the eyes of an ordinary reader. The same is true of artists and other producers of instructional materials. In order to get a fresh appraisal of the draft material, you can ask other people to comment on it, preferably people who are well qualified to assess this particular type of material.

You might consult different experts for different reasons. If it was a draft lesson from a correspondence course in law, you might send it to a lawyer to check the content of the lesson, but also to a schoolteacher or to an expert in course writing to assess the instructional quality. Try to list the aspects of the material that you want to have comment on, and select an appropriate expert (or perhaps several) for each aspect.

Several experts are better than one. An expert's comments reflect his particular experience and the general opinions that he already holds; two experts can give completely different comments, even contradictory ones, about the same material.

You have to decide whether to consult the experts individually or to gather them together to give their comments as a group. They are more likely to find time to read the material if they have to discuss it in a group, and the group discussion may give rise to points that none of the participants would have thought of on his own. On the other hand, people in a group tend to defer to the most respected or the most forceful members; someone who was going to say something different is likely to keep it to himself, not wanting to appear foolish or to provoke an argument. Comments given individually by experts may be more varied than the opinions expressed by a group.

When you submit some draft material for expert scrutiny, the comments that you least expect are often the most useful. The outsider, bringing a fresh pair of eyes to the material, can draw attention to things that the materials producer had not considered. For this reason, it is better to make an open request for comment, rather than preparing a questionnaire for the experts to answer, though you may sometimes want to direct their attention to certain aspects of the material, to make sure they they comment on those aspects, whatever else they may say.

While the criticisms of experts can cause the materials producer to make useful revisions, expert judgements are not a substitute for pre-testing. The material's real job is not to please the experts but to communicate to the students. Experts can give opinions on whether the material will communicate successfully or not, but they are not clairvoyant. The only way to find out for sure is to test it.

Measuring readability of English text

Some books and articles are easier to read than others. This is partly because of the content; you would expect an article on nuclear physics to be more difficult than an article on football. But it is also because of the style. There are many books which explain basic statistics, for example, but some are easier to read than others; all the books explain the same things, but some explain them better. The word 'readable' is often used to describe a book that is easy to read. In this context, 'readable' does not necessarily mean that the book was lively, interesting or pleasant to read, only that it was easy to read. That's how I will use the word in this section.

If you are writing instructional materials, you obviously want to make them as readable as possible, especially if the language they are written in is the students second language. As part of the preliminary assessment of some material, such as a few draft lessons of a correspondence course, it might be useful to make an estimate of the level of readability. The writer has tried to write in a clear and simple style, but is it clear and simple enough for the students? How can you tell?

The best way to tell is to give the draft lessons to a few typical students and see how well they can understand them. However, if the material is in English, there is something else you can do before you go to the trouble of getting some students to read the lessons; you can make a rough estimate of the level of readability by calculating a readability score. If the score indicates that the draft is much too difficult, then you can say that the lessons need to be rewritten before doing any pre-testing.

Most people, especially experienced writers and editors, can make a good judgement of how readable a passage is. However, their judgements often differ; the writer might think his draft is very readable while the editor thinks it's too difficult. In order to improve on personal judgements, English-language researchers have made great efforts to devise an accurate way of calculating how readable a passage is. They have produced dozens of 'readability formulas'. Though these vary in detail, they share the same basic idea, so I will describe just one of them. This is a simplified version of the Flesch Reading Ease score devised by Farr, Jenkins and Paterson. I have chosen it because it is one of the simplest to calculate.

This method is based on the finding that passages in English that are hard to read tend to contain a lot of long words and long sentences. This is not to say that all long words are hard or that all long sentences are hard; it's just that these things tend to go together. The Reading Ease score (abbreviated to RE score) is a measure of the proportion of long words and long sentences in the passage. If the RE score shows that a passage has a lot of long words and long sentences in it, this suggests that the passage is difficult to read. It does not prove that the passage is difficult to read, but it suggests that it probably is.

The RE score is no more than an indicator; it indicates the readability of a passage rather in the way that the price of a camera indicates the quality of the camera. In general, the price is a fairly reliable guide to the quality of the camera; the more expensive cameras are generally better than the cheaper ones. But the price is occasionally misleading; a mediocre camera can be over-priced, or a fairly cheap camera might be unexpectedly good. So it is with the RE score. As a rough guide, it is generally reliable, but it can be misleading. You should use it alongside personal judgements of a passage.

To calculate the RE score, for a book or an article or a correspondence lesson, you proceed as follows:

1. Take five passages from different parts of the piece, each about 100 words long. Begin each extract at the start of a paragraph. End it at the end of a sentence. Count contractions as one word, e.g. 'He's'. Count hyphenated words as one word, e.g. 'book-keeper'. Count numbers as words, e.g. '1914' as one word, '7, 8' as two.

2. Count the number of monosyllables (words of one syllable) in each passage. Whether a word is a monosyllable depends on how it is pronounced, not on how it is spelt. For example, 'any' has two syllables even though it only has three letters, whereas 'strength' has only one syllable even though it has eight letters. For symbols and figures, count them as they are read aloud, e.g. '10' is a monosyllable, '11' is not; '£' for pound is a monosyllable, '$' for dollar is not.

3. Count the number of sentences in each passage. Where sentences are separated by a colon (:) or semi-colon (;) count them as separate sentences.

4. Add up the total number of words in the five passages, then the total number of monosyllables, then the total number of sentences. The following is for five passages from a maths correspondence course produced by LDTC:

Passage number	Words	Monosyllables	Sentences
1	100	66	9
2	98	65	7
3	106	67	7
4	90	63	8
5	110	71	5
Total	504	332	36

5. Divide the number of monosyllables by the number of words, and multiply by 100. In the example above,

$$\frac{332}{504} \times 100 = 66$$

This gives you the number of monosyllables per 100 words.

6. Divide the number of words by the number of sentences.

$$\frac{504}{36} = 14$$

This gives you the average sentence length.

7. Refer to the table in Appendix 7 for the Reading Ease score. In the above example you look down the column headed 66 and along the row headed 14, and you find the RE score: 60.

If you calculated the RE score on the basis of the entire book, instead of on just five passages, you would probably find that the score was within 5 points of this one. That is, the RE score for the maths course in the above example is between 55 and 65.

The RE score goes from 0 to 100. A score of 0 means that the passage is almost unreadable; a score of 100 means that the passage is extremely simple.

Here are two examples taken from correspondence courses. Both courses were about business studies, but they were written by different colleges and at different levels - the first at an elementary level, the second more advanced.

> 'Mr Smith buys goods from wholesalers and sells these goods to his customers. This is how he makes money. Mr Smith also makes money from other things. For example, he has a small van of his own. Sometimes he transports goods for his customers. Every time he carries goods for his customers Mr Smith makes them pay for this service. Book-keepers call this "Transport Receivable" because Mr Smith receives money. He does not sell the van. He gives a service and receives money. He sells a service.'

> 'The degree to which production control is developed varies with different industries. It is at a minimum where a single homogeneous product is treated by a fixed sequence of processes in a continuous flow. Modern examples on a vast scale are afforded in paper, pulp and petrochemical industries. Flow charts in these industries exhibit a continuous stream of production in which many operations are performed, materials added, and by-products or wastes eliminated, but without break in flow or exceptions in work or processes.'

The course from which the first quotation was taken had an RE score of about 63; the second course had an RE score of about 37. Our experience at LDTC suggested that, when writing materials in English for people who were studying for the Junior Certificate exam (which schoolchildren take after three years of secondary school), we should aim for an RE score of about 60 to 75, i.e. more like the style of the first quotation than the second. If a draft lesson scored below 50, the writer and editor should have another look at it.

Two words of warning about the use of the RE score. The first is that it was standardised on American prose and for American readers. It seems fairly safe to assume that if Americans would find a passage difficult, then second-language readers of English would also find it difficult. But it is not necessarily true that if Americans would find it easy, second-language readers would also find it easy. Second-language readers have notorious difficulty with phrasal verbs - these are expressions in which a simple verb is coupled with a preposition to give it a special meaning, such as 'give up', 'give out', 'give in'. Most phrasal verbs consist of monosyllables, so a passage with many phrasal verbs would have a high RE score, but second-language readers could still find it difficult. For second-language readers, for instance, one should use 'tolerate the cold' rather than 'put up with the cold' and 'postpone the ceremony' rather than 'put off the ceremony'.

Secondly, if you decide to redraft a passage because the RE score suggests it is too difficult, you should start again from scratch and think how you could express yourself more simply. It is no good at all merely to tinker with a few words and sentences in order to raise the RE score. You do not make the passage appreciably more readable just by changing a few words and chopping a few sentences in half, though this might increase the RE score dramatically. To return to the example of the camera prices, the price is a rough guide to the quality of the camera, but a salesman obviously does not improve the quality of a camera simply by increasing the price. Likewise, the RE score is a rough indication of the readability of passages that have already been written, but you do not increase the readability of a passage just by artificially raising the RE score.

The RE score can be used only for text in English. Passages with a lot of long words, in English, are generally more difficult to read, and this is one of the things that the RE score depends on. But this might not be true of other languages. If you calculated the RE score for a passage in, say, Sesotho, you could get a very misleading result. Readability measures have been devised for Spanish (see Appendix 5), though they are not as easy to use as the RE score.

If you were producing materials in a country's local language, you would probably not be able to do a readability measurement of this kind in advance of pre-testing. However, if you wrote a passage in English, for subsequent translation into a local language, it would still be worthwhile testing the English version by calculating the RE score and improving the the passage if necessary, on the grounds that a translation of a simple passage of English is likely to be clearer than a translation of a difficult passage of English.

13 Pre-testing instructional materials

This chapter describes ways of testing instructional materials to check that they convey what they are supposed to convey.

Watching people use the material *Simply watching people read some material, work through exercises or listen to a programme can indicate the parts that may give trouble.*

Assessing comprehension *Asking people questions will show if they have interpreted a picture or sound in the way they were meant to.*

Testing the material, not grading the students *The sort of comprehension test that schoolteachers write for their pupils is not the sort you want for pre-testing materials.*

Writing test questions *Short-answer, two-choice and multiple-choice questions.*

Analysing pre-test results *Ways of tabulating the results to see which parts of the material need improving.*

Cloze testing *This is a way of testing a passage, in any language, to find out how difficult it would be for people to read it.*

Getting reactions to the material *Ways of finding out people's feelings about the material.*

Most people acknowledge the value of pre-testing instructional materials, as a general principle, but they are tempted to ignore it when it concerns actual pieces of material that they themselves have produced. If a writer, an artist, or a radio producer has been working on some material, he is naturally very familiar with it and he tends to assume that the material is just as clear to everyone else as it is to him, so he is inclined to think that pre-testing his particular piece of work is unnecessary. In addition, the work is probably behind schedule; people are anxious to meet production deadlines and they look upon pre-testing as a needless delay.

You can make this problem less severe by making sure that time is set aside for pre-testing in the production timetable. Pre-testing should be regarded as a routine stage through which all materials must go, rather like the proofreading of typescripts. Secondly, you can use pre-testing methods that produce results quickly; most of the methods I will describe in this chapter take only one or two days. Even so, the pressure to skip the pre-testing can be quite strong, and the researcher must be prepared to resist it. In my experience at LDTC, pre-testing always resulted in some appreciable improvement being made to the material and occasionally showed up a serious flaw which, if left uncorrected, would have been disastrous.

Watching people use the material

A simple way to pre-test a correspondence lesson is to gather a group of students together, give them copies of the draft lesson and watch them while they work through it. A colleague in Swaziland pre-tested some lessons of an English course in this way. In the draft lessons he had included a set of test questions at the end of each one. By watching some students work through these lessons, he discovered that several of them turned to the test questions first and answered them before they read the lesson. In the revised version he included a clearer explanation of the purpose of the test questions.

The best person to carry out this pre-testing is, of course, the writer himself. If he is there, the students can also ask him questions if they have any difficulties. This will reveal the parts that give difficulty and their questions will also indicate what has caused the trouble.

You can get more out of this exercise by giving the students a coloured pencil and asking them to underline the parts they find difficult. Better still, you can give them two coloured pencils and ask them to underline easy parts in one colour and hard parts in the other. People are sometimes reluctant to underline only the hard parts, as this might suggest that they are critical of the writer or that they themselves are not sufficiently intelligent to understand the text; they probably feel happier if they are asked to underline the easy parts as well.

After the students have worked through the lesson, a group discussion might show up further problems; the students might feel that the lesson does not go into enough detail, for example, or that the style is too childish.

Similar techniques can be used to pre-test a leaflet for rural people. You take a draft of the leaflet out to a village and ask some people to read it. One important thing to note here is the length of time that people take to read the draft; people with limited reading ability take much longer than you might imagine to read even a short leaflet. If the people don't mind, you can ask them to read the leaflet out aloud; the way they read it will indicate any difficulties they have with certain words or sentences. Here again, a group discussion can be helpful.

Assessing comprehension

Sometimes instructional materials are designed to teach people how to do things, such as how to crochet a baby's bonnet, or how to make a vegetable grater out of a tin can. The best way to pre-test such materials is simply to see if people can follow the instructions. At LDTC we pre-tested a crochet booklet by giving draft copies to a few housewives with balls of wool and crochet hooks, and asking them to teach themselves to crochet.

Generally, however, instructional material is designed to inform rather than to give instructions on how to perform a particular task - a radio programme for farmers about conserving pasture, for instance, or a lesson on rock formations from a correspondence course in geography.

The most natural way to see if people have understood it correctly is to ask them some questions about it.

With pictures and sound effects, one simply wants to know if people can recognise them in the way that was intended. You show them the pictures, or play them the sound, and ask them 'What is it?' If possible, you should show them the picture or the sound effect in its context. The sound of a car crash might be difficult to recognise on its own, but easy to recognise when played as part of a short spot about road safety.

You have to pre-test materials on the sort of people for whom they are intended. Say a nutrition poster has pictures of milk bottles and cabbages. It is not enough just to find out that people in the office can interpret these pictures correctly. If the nutrition poster is intended to be seen and understood by rural mothers, you have to make sure that rural mothers can interpret it correctly. This kind of pre-test can be regarded as a small social survey and, though you would not use an elaborate sampling technique, you still want your sample to be representative. See Chapter 4.

You might have to ask a few more questions in addition to 'What is it?', in order to find out if people understand fully the picture or the sound. We tested a drawing (Fig. 1) which had been used on the cover of a family planning pamphlet. Although everyone recognised that it was a man and a woman, half the people answered 'No' to the question 'Do you think they are Basotho?' This was a serious fault in the picture since the readers, who were mainly rural Basotho, were intended to think that the pamphlet was about people like themselves, not about foreigners.

Another example is shown in Figure 2. Readers were intended to think that this was a couple with a child, that they were happy and that they were fairly wealthy. The artist had indicated that the couple were wealthy by putting them on a padded sofa. In the pre-test, we found that everyone saw that it was a couple with a child, but when we asked, 'Are they happy?' several readers said 'No'. We asked them why they said this and they replied, 'Because the couple are sitting in a thorn bush'; the lines that the artist had drawn to represent creases in the buttoned cloth were misinterpreted as thorns.

If the material is a pamphlet or a correspondence lesson or a radio programme, you want to know if people can understand it and learn from it. You let the people read the pamphlet or the lesson, or you let them listen to the programme, and then you test their understanding of it. (With correspondence students you might send the draft lesson and the test by post.) With students or schoolchildren, you can administer a pencil-and-paper test, but with rural people it is best to administer the test as an interview, i.e. the interviewer reads out the questions and the respondent gives the answers. This is because many rural people, even if they are literate, are not accustomed to taking pencil-and-paper tests so their results on a written test would indicate how good they were at taking written tests rather than how well they had understood the instructional material.

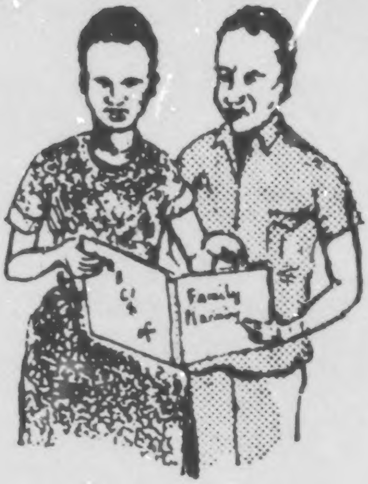


Figure 1

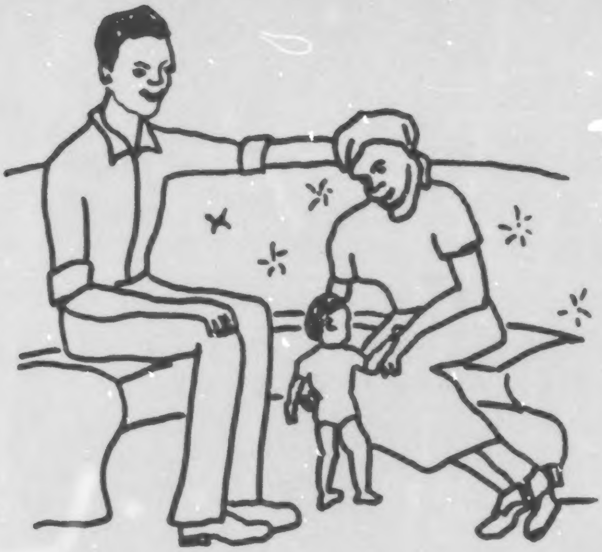


Figure 2

Testing the material, not grading the students

When a researcher sits down to write questions that will test whether people have understood some material, he is inclined to think back to his schooldays and to write the sort of comprehension test that he remembers from then. This is unfortunate because he is likely to produce the wrong sort of test for pre-testing material.

In the first place, when a teacher gives his pupils a passage to read and then gives them a comprehension test on it, the test is directed at the pupils rather than at the passage. He is not particularly concerned about the actual passage that he uses in the test; any one of a large number of passages would have done just as well. The purpose of the test is to assess how well the pupils can perform on this kind of task. He is testing the pupils' ability to comprehend the passage rather than the passage's effectiveness in communicating to the pupils.

Secondly, a lot of the tests that a teacher gives to his pupils are designed to rank the pupils in order of ability. He wants a test which will show him which of his pupils are good at comprehension, which are average and which are poor. A test which was so difficult that every pupil got no marks would be no use to him. Likewise, a test which was so easy that every pupil got full marks would be no use. He wants a test where the clever pupils will get quite high marks, the average pupils will get middling marks and the poor pupils will get low marks.

Consequently, the sort of test that a schoolteacher tries to write is one that most of his pupils will find quite hard. It will contain some questions of middling difficulty to sort out the average pupils from the poor ones (i.e. an average pupil will get them right but a poor one won't), and some really tricky questions to sort out the very good from the fairly good.

Now, when pre-testing some material, the purpose of the test is obviously different from this. You are not the slightest bit interested in whether Mr Motsamai, who happens to be the first person you interview, does any better or worse on the test than Mrs Nkekele, who happens to be the second. You are not going to list the people who take the test in order of achievement. You just want to know whether they can understand the material. So, a test that a researcher writes to pre-test some material is going to be different from the sort of comprehension test that a schoolteacher would write to test his pupils.

You start by asking the writer to list the main points that he wanted to get across, or you read the material and make your own decision about what seem to be the main points. (If the writer has specified behavioural objectives - see Chapter 12 - you make use of them at this stage, of course.) And then you write questions to find out if the readers have understood the main points.

It does not matter if the questions seem very easy. In fact, if the material is clear and simple and communicates effectively, and if the questions are straightforward, it ought to be easy to answer them. If the questions seem hard, there is probably something wrong with them, or with the material. A schoolteacher might sit back in satisfaction after writing a test and say to himself, 'That will sort them out.' Sorting out the pupils is the purpose of that kind of test. But a researcher ought not to feel like that about a test he's written for pre-testing some material.

While the questions should be easy for someone who has read and understood the material, the answers should not be obvious even to someone who has not read the material. For example, you might be pre-testing a pamphlet about fertilisers. It would not be much use to ask, 'Do fertilisers help plants to grow?' Most people could answer correctly without reading the pamphlet. On the other hand, you should not concentrate on minor details. Decide what are the important points in the material and then ask questions to find out if the people have taken in these points. If the material has been successful, then most of the people will get most of the answers right.

Writing test questions

When writing questions for a pre-test, it is tempting to choose open questions such as, 'What do you think are the main points in this pamphlet?' It does not require much effort to think up such questions. If you are pre-testing the material on only four or five people, you can get away with questions of this kind. In fact, they can produce interesting results - perhaps readers of the pamphlet draw different conclusions from what the writer intended. However, if you collect answers from more than a handful of people, you encounter the main problem with open questions, which is that the answers are difficult to record and analyse. It might be easy to think up the questions but it's hard to process the answers.

To avoid this problem, you should try to write questions which require short answers. Such questions are harder to write but the answers are much easier to mark.

When administering the test in an interview, the best questions are those that require the respondent to give a particular word or phrase as the answer (known as 'short-answer questions'), such as 'What fertiliser does the pamphlet recommend for beans?' or 'If you apply Pestex to a vegetable garden, how long should you wait before harvesting the vegetables?' Try to keep the questions as short and as clear as possible. Obviously, if a question is long and complicated, people will have difficulty with it, even though they might have understood the material perfectly. In particular, avoid negatives; you would create needless confusion by putting a question in the form, 'What pesticides are not recommended for use against cutworm?' Of course you can use a negative if there is a special reason for it. For example, if a pamphlet on first aid stresses that you should not use greasy ointment to treat burns, you might ask 'According to the pamphlet, what should you not use when treating burns?' With written material, you should allow the respondent to keep it and refer to it, if he wants to, while answering the questions; you're testing his comprehension, not his memory.

Another type of question that you can use in an interview is the two-choice type, where the question allows only two possible answers, such as 'Yes, No' or 'True, False'. These are not as good as the short-answer kind because the respondent has a 50% chance of getting the right answer just by guessing. Suppose you ask 'Does the pamphlet recommend the use of horse manure?' If the respondent says 'Yes', and 'Yes' happens to be the right answer, it does not necessarily show that he has read and understood that part of the pamphlet. However, there are times when it is difficult to frame a question in short-answer form, so you have to fall back on the two-choice type.

As with short-answer questions, keep two-choice questions short and clear. Make sure that the correct answer is indeed one of the two available and not something in between: it would be wrong to ask, 'Should you make a child vomit if he's swallowed something poisonous?' if the answer was 'Often, but not always' or 'It depends on what he's swallowed'. If you have several questions of the two-choice type, don't have the same correct answer for all of them, e.g. all 'Yes'. And don't make the 'Yes' ones visibly different from the 'No' ones, for example by making the 'Yes' ones longer than the 'No' ones.

If you're administering a pencil-and-paper test, rather than an interview, there is another type of question you can use - the 'multiple-choice question'. Here are two examples:

What pesticide is recommended for use against cutworm?
A Pestex
B Thiodan
C Dieldrin
D Paraquat
E Maladrin

If someone has scalded his hand with boiling water, the first thing you should do is:

A put his hand into cold water
B make him sit down
C wrap him in a blanket
D wrap the hand in a clean bandage

The first part - a question in the first example and an incomplete sentence in the second - is called the 'stem'. This is followed by four or five alternative answers, only one of which is correct. The other answers, which are called 'distractors', are either wrong or not as good as the correct answer.

The multiple choice question is better than the two-choice type in that there is less chance of getting the right answer just by guessing - only a 20% or 25% chance, in fact, depending on the number of alternatives. With well chosen distractors, it can also indicate which errors people are likely to make, and this can give you ideas about how the material needs changing. And, of course, the marking is very quick. The main disadvantage is that they are difficult to write; thinking up plausible distractors is particularly hard.

When writing a multiple-choice question, keep the alternatives as short as possible; the stem should present the problem and the alternatives present possible solutions to that problem. Try to think up distractors which are plausible. The following question would be too easy because the distractors are not sufficiently plausible:

If someone has scalded his hand with boiling water, the first thing you should do is:

A put his hand into cold water
B give him a glass of brandy
C tell him a joke to take his mind off the pain
D make him a strong cup of tea

Don't give unintentional clues to the correct answer. Suppose you were pre-testing a correspondence lesson with a set of multiple-choice questions. If you had unintentionally written it so that the correct answer was always the longest, a perceptive student could score well on the test without having understood the lesson at all. The following example contains an obvious grammatical clue to the right answer:

The longest side of a right-angled triangle is called:

A invented by Pythagoras
B the hypotenuse
C opposite the right-angle
D the sum of the other two

If you have a series of multiple-choice questions, vary the position of the right answer in a random way, i.e. write all your alternatives for a question and then use some randomising device, such as a dice or a table of random numbers, to decide which should be A, which should be B, and so on. Novice test writers tend to avoid putting the right answer at position A or E, and this, of course, gives an extra unintentional clue to the student.

While you should be careful not to give unintentional clues, you should also avoid complications that make the questions needlessly difficult, especially if the students are taking the test in their second language. The question should be sufficiently clear so that someone who has the required knowledge can give the right answer. Avoid using negatives, or implied negatives, unless there is a special reason for it. In the following example, someone who knew the right answer ('Acute') could be misled by the confusing expression 'fails to exceed' into giving a wrong answer. If you used this question as part of a pre-test of a correspondence lesson in geometry, the results would not indicate that the students had failed to learn about acute angles, only that they found this question hard to understand.

What is the name for an angle that fails to exceed 90 degrees?

A Obtuse
B Scalene
C Acute
D Reflex
E Polygon

For the same reason, you should also avoid using the phrases 'none of these' and 'all of these' among the alternatives. Consider these four sets of alternatives for the same question:

Which figure contains, by definition, a pair of parallel sides?

1. A Pentagon
   B Quadrilateral
   C Kite
   D Trapezium

2. A Pentagon
   B Quadrilateral
   C Kite
   D None of these

3. A Trapezium
   B Rhombus
   C Parallelogram
   D All of these

4. A Trapezium
   B Quadrilateral
   C Kite
   D All of these

The first set are straightforward; if the students know that a trapezium has a pair of parallel sides, they've got the answer(D). But they need to know more, and to think harder, to get the right answers for the others (D, D and A respectively). If they picked wrong answers from the first set, you'd know they hadn't learnt about the trapezium. But if they picked wrong answers from the other sets, you couldn't be sure exactly what it was that they didn't know.

Whatever sorts of question you use in a paper-and-pencil test, make sure you give clear instructions to the students about how to record their answers. Of course it will seem obvious to you, but it will not be equally obvious to the students. So say exactly what they have to do, perhaps illustrating it with an example. And remember to say how they can cross out an answer and change it if they want to.

Analysing pre-test results

If a materials producer has specified what he will take as an acceptable level of performance from people who have studied his material, the first thing to do is to mark people's answers and count them up to see if they've done satisfactorily. Say a course writer has said that he expects every student to get at least three-quarters of the answers right. If every student has scored above 75%, then the lesson is generally satisfactory, though perhaps some details could still be improved. If any is below 75%, there's something wrong and the course writer will have to change the lesson.*

Pre-test results, handled in this way, are acting as traffic lights; a satisfactory result means you can proceed to the next stage, whereas an unsatisfactory one means you should stop and do something about it. But pre-test results can play a more useful role in showing you precisely what aspects of the material need to be changed.

If you asked open questions in the pre-test, such as, 'What is this a picture of?' you should look through the various answers that people gave. If the same answer crops up several times, you can tabulate the results. Say you had pre-tested a drawing of a cabbage for a nutrition poster by asking thirty people, 'What is this a picture of?' You might tabulate the results as follows:

People who said it was a		cabbage	14
"	"	tree	8
"	"	bird	3
"	"	hat	1
"	'don't know'		4

This is more useful to the artist than simply saying 'About half the people got it wrong' since it might suggest to him what it was that was misleading people and how he could improve the drawing.

---

* He might have specified his objective in a more complicated way, saying for example, that half the students should get above 60% and even the worst student should get at least 40%.

If the pre-test contains a set of questions, it is useful to look at the results separately for each question. (This is known as an 'item analysis'.) You generally have only a small number of respondents for a pre-test - say under 40 - so you can used the squared paper method for this (see Chapter 7). You put the question numbers across the top and the respondents down the side. The following shows how a squared sheet might look if you had given ten questions to twelve people, coding their answers as Right - 1, Wrong -2, Don't know - 3, No reply - 0.

		Question									
		1	2	3	4	5	6	7	8	9	10
Respondent	01	1	1	3	1	2	2	1	1	0	1
	02	1	2	1	1	2	0	1	3	1	1
	03	1	1	1	1	2	3	1	1	1	1
	04	1	0	3	3	3	3	3	3	3	3
	05	1	1	1	1	1	2	1	1	1	1
	06	1	1	3	1	2	3	1	1	2	1
	07	1	1	2	1	3	3	0	0	0	0
	08	1	1	1	1	2	1	1	1	3	1
	09	1	1	3	1	2	2	1	2	2	1
	10	1	1	2	1	2	3	1	1	0	1
	11	1	1	1	1	2	2	1	1	1	1
	12	1	1	3	1	2	3	1	3	0	0
Total right		12	10	5	11	1	1	10	7	4	9
Total wrong			1	2		9	4		1	2	
Total 'don't know'				5	1	2	6	1	3	2	1
Total no reply			1				1	1	1	4	2

By looking at the columns, you can see at a glance which questions gave the most difficulty. In this example, questions 5 and 6 were poorly answered, followed by questions 3 and 9.

If your questions are in multiple-choice form, you can analyse the results in even more detail by seeing which distractors the students choose. The results for one question from the pre-test of a science lesson might be tabulated as follows:

If you put one end of a metal poker in a fire, the other end gets hot after a while. The process by which heat moves along the poker is called:	Number of students who chose this answer
A combustion	3
B radiation	0
C induction	2
D convection	11
E conduction	17

Although most of the students got the right answer, the teacher could infer that many were unclear about the difference between conduction and convection, so he should try to clarify this point in redrafting the lesson.

Cloze testing

At LDTC, I once collected a set of passages in Sesotho - some from primary school reading books, some from LDTC's own publications and some from newspapers and novels - and I gave these to about a dozen of my Basotho colleagues. I asked each person to make his own assessment, without consulting other people, of the readability of each passage and to give a readability score to each passage, from 0 for very difficult to 10 for very easy.* To my surprise, the scores that my colleagues gave to these passages varied greatly from one person to another. In fact, the variation was so great that, in some cases, one person thought that a certain passage was the easiest and another thought it was the hardest.

I am fairly sure that they all understood what I wanted them to do, and I checked to make sure that no-one had got his scoring upside down (0 for very easy and 10 for very difficult). I never found a satisfactory explanation. Perhaps there are several different aspects of a passage in Sesotho which you could take as a criterion of readability, and different people attended to different aspects. Or perhaps the passages I had selected were at the same level of readability, despite their apparent variety, so my colleagues gave their scores on the basis of some other characteristics of the passages. (My own command of Sesotho was not good enough for me to form a judgement myself.)

---

* As on pages 167 to 171, I am using the word 'readable' to mean 'easy to read'.

The point of this story is that, if you ask people to assess the readability of a passage (a draft pamphlet on farming, for example), one person might tell you that it is clear and simple and another person might tell you that it isn't. If you are fluent in the language yourself, you will be inclined to trust your own judgement, of course, but this does not really solve the problem. What you want is some way of measuring, objectively, how readable a passage is. There is a way to do this; it's called 'cloze testing'. But before I describe it, I should admit that it requires quite a lot of work. For this reason, although we experimented with it at LDTC, it never became a routine pre-testing procedure.

Cloze testing was invented for English but it has since been shown that you can apply it to other languages, even to ones that are linguistically unrelated to English. The person who invented it borrowed the term 'cloze' from a branch of psychology. The word is pronounced exactly like the ordinary word 'close' as in 'Please close the window'.

To do a cloze test on a passage, you type out the passage leaving out every fifth word, putting in a line about 2 cm long for every word you leave out. So it looks like this:

> In traditional farming, leaves ________ branches are burned, the ________ destroys the organic matter. ________ quantity of organic matter, ________ provide humus, is thus ________. When cassava is harvested, ________ whole plant is lifted; ________ roots to eat, the ________ to be replanted. Almost ________ organic matter is left ________ the soil or in ________ soil. The cassava has ________ humus from the soil, ________ the organic matter of ________ cassava has not been ________ to the soil.

Be sure to leave out every fifth word, whatever the word is, even if the word is 'the' or 'is' or 'and' or 'Fred'. Treat numbers as words, e.g. '1946' is a word in 'He was born in 1946.'*

You then make 20 to 30 copies of this and give them to people. You ask them to write in what they think the missing words are. If they have no idea what the missing words are, they can leave a blank, but you should encourage them to make a guess if they can. Collect the papers in when they have finished. Give a mark for every word they have guessed correctly. Ignore minor spelling mistakes, e.g. if the person has written 'leeves' when the correct word was 'leaves', count that as correct. But do not count synonyms, e.g. if the correct word was 'leaves' do not accept 'foliage' even though you might feel that 'foliage' would have been just as good. Count up all the correct guesses and divide by the total number of blanks and multiply by 100. This is the 'cloze score' for the passage.

---

* There may sometimes be difficulty in deciding what counts as a word, and this problem will arise more in some languages than in others. It doesn't much matter what decision you make so long as you are consistent, i.e. you must treat similar cases in the same way whenever you prepare a passage in that language for a cloze test.

For example, say the passage contained 25 blanks and you gave it to 30 people. Altogether, therefore, they had 25 x 30 blanks = 750. Let's say that they made 204 correct guesses, altogether. The cloze score for the passage is $\frac{204}{750} \times 100 = 27$.

You can make the test more reliable by producing two test versions of the passage, instead of one. In the first, you blank out the fifth word, then the tenth, then the fifteenth and so on. In the second version, you blank out the fourth word, then the ninth, then the fourteenth and so on. Using the same example as above, the first two lines of the second version would look like this:

In traditional farming, ________ and branches are burned, ________
fire destroys the organic ________. The quantity of organic

If you produced two versions, you would give one version to half the people and the other version to the other half. In working out the cloze score for the passage you add all their correct guesses together (i.e. you ignore the fact that Jane and her friends had one version while Mary and her friends had the other; you just add their scores together.)

If you have the time and the patience, you can produce five different versions of the same passage, giving each version to one fifth of the people. The third, fourth and fifth versions of my example would start like this:

In traditional ________, leaves and branches are ________,
the fire destroys the ________ matter. The quantity of ________

In ________ farming, leaves and branches ________ burned,
the fire destroys ________ organic matter. The quantity ________

________ traditional farming, leaves and ________ are
burned, the fire ________ the organic matter. The ________ of

To get a reliable cloze score for a passage, you should give it to at least twenty people, and preferably more.

What does the cloze score tell you? The higher the score (i.e. the more correct guesses people made), the more readable the passage. A high cloze score means that people would find the passage easy to read; a low score means that they would find it difficult. If you have only one cloze score - say 27 - it does not tell you anything since you do not know whether 27 is a high score or a low one. You have to do cloze tests on several different passages (in the same language*) on the same people, or at least the same sort of people, before you can judge whether a particular score is high or low.

---

* You can't compare cloze scores for passages in different languages.

Say you had tested five passages on schoolchildren in standard 6, for example, and you had obtained scores of 13, 19, 21, 25 and 29. If you then tested a new passage on them and got a score of 27, you would guess that 27 was a high score; so you could conclude that this new passage was quite readable compared with the previous five passages.

If you used this test regularly, you would get a good idea of the range of scores to expect, and you could say with more confidence whether a particular score was high, or low, or about average. In order to do this, you would have to do all your testing on the same sort of people, such as first-year secondary schoolchildren. This is because more educated people get higher scores than less educated ones. If you tested one passage on primary schoolchildren and got a score of 14, and another passage on secondary schoolchildren and got a score of 24, you still would not know which passage was more readable; you would expect secondary schoolchildren to get a higher score, whatever the passage.

Like the RE score for English (Chapter 12), the cloze test is not infallible. It should be used alongside personal judgements. However, as I found at LDTC, people can make very different judgements about the readability of passages. Where one person thinks a passage is easy and another thinks it is difficult, the cloze test can provide a useful extra measurement.

Getting reactions to the material

Most of this chapter has been about ways of testing whether people can understand and learn from the material. That is usually the most important thing. But you might also want to see whether they find the material interesting, or boring, or in some way offensive.

There is a form of attitude measurement known as a 'Likert scale'. You present a series of statements such as, 'The lesson was too long', 'The lesson covered the topic thoroughly', 'The pamphlet tells you all you need to know', 'The pamphlet is suitable for young people to read', and so on. Next to each one you provide a scale such as 'Strongly agree, Agree, Disagree, Strongly disagree'. So your questionnaire might look like this:

	Strongly agree	Agree	Disagree	Strongly disagree
The lesson was easy	-	-	-	-
The lesson was dull	-	-	-	-
The lesson prepares you adequately for the test at the end	-	-	-	-

In an interview, the interviewer reads out the statements, after explaining that the respondent has to choose one of these four alternative answers for each one. With a self-completed questionnaire, the respondent simply puts a tick to indicate his chosen response.

One of the big problems in getting people's reactions to draft material is that people are polite. If you give a draft correspondence lesson to some students and then ask, 'Did you find it interesting?', most of them will say Yes even if they actually found it rather boring. For this reason, you can't take the results of straight attitude questions at their face value. 'Ninety per cent of the students found the lesson enjoyable,' does not necessarily mean that the lesson was particularly enjoyable. However, such results can be useful if you are testing two different versions of a lesson. Suppose you divided a class of students randomly into two groups and gave version A to one group and version B to the other; 90% of those with version A said they enjoyed it, as against only 60% of those with version B. Perhaps both figures overstate the amount of enjoyment the students got, but you can still conclude that version B was less enjoyable than version A.

One way to overcome this obstacle of politeness is simply to instruct people to be critical, e.g. 'List three things that you liked and three things that you did not like about this lesson.' Another way is to ask them to compare one part with another, e.g. 'Which part did you find the most difficult?'

With materials such as pamphlets or radio programmes, the best way to judge people's reactions is to watch them. It will be fairly obvious if they find a radio programme boring, or a pamphlet hard to read. A discussion afterwards might bring out further points. For example, in pre-testing material on family planning, one might be concerned about its acceptability. One way to prompt discussion of this would be to put a hypothetical question, such as 'If this programme was being broadcast, would you be happy for your children to hear it?'

A final word of caution about relying on people's reactions to draft material. If people react badly to it, then there is certainly something wrong with it and it needs changing. But if people express approval of it, it does not necessarily mean that the material is satisfactory. This is partly because of the politeness problem that I have already mentioned. But it is also because people often have nothing to compare it with.

Suppose you had a draft pamphlet on immunisation of children and you gave it to a rural mother to get her opinions on it. By what standard is she to judge it? The questions in the mind of the pamphlet writer are 'Is it as good as other pamphlets?' 'Is it good enough for the job we want it to do?' 'Is it as good as we can make it?' But the question in the mind of this rural housewife will be 'Is it better than having no pamphlet at all?' And her answer, almost certainly, will be 'Yes'. If she expresses approval of it, she probably means that she is in favour of immunisation and that she thinks people should be told about it and so this pamphlet is a good thing. Her approval of the pamphlet does not guarantee that there is nothing wrong with it.

14 Keeping regular records (monitoring)

The sort of records you keep depends very much on the way your organisation works. In this chapter I describe the sort of system that a small correspondence college might use.

Who are the students? Enrolment forms and record cards can be used to provide an annual profile of the students.

How are the students progressing? Regularly updated record cards - one for each student - can form the basis of periodic progress reports.

Which parts of the courses need improving? Regular reports on the courses are useful when the time comes to update them.

Exam results Exams provide feedback on the college's performance but the results have to be interpreted carefully.

There are two good reasons for collecting information at regular intervals on the activities of a distance-teaching organisation. The first is that it enables you to monitor the system. This means keeping an eye on the system so that you can see if things are not going satisfactorily and do something about it. The word 'feedback' is sometimes used to describe this process. The idea is that the distance-teaching organisation puts out its materials (correspondence lessons, radio programmes or whatever), and information is fed back to the organisation on how the materials are being received. The second reason for collecting information regularly is that it provides a basis for evaluating innovations. For example, if you are broadcasting a weekly radio programme and you have collected information at regular intervals on the audience, you can see whether a change in the programme makes any difference to the audience.

Sometimes you have to set up a special system for collecting the information you want. If you wanted to have records of the quality of radio reception around the country, for example, you might arrange for a number of listeners to keep a note of the reception, according to some agreed system, and to send in monthly reports. Sometimes, however, you can make use of records that are already collected for administrative purposes, perhaps arranging with the administration or accounts departments to have the records collected in a way more helpful to you. If your organisation was selling instructional booklets, for example, there would be some system for recording the sales; perhaps you could arrange to use these records to observe total sales per month or to compare sales in different parts of the country. By co-operating with other departments you may be able to collect a lot of useful information without putting anyone to much extra trouble or expense.

When designing a monitoring system that produces regular reports for someone, you should bear in mind the range of actions that the person in question can take - the student adviser, the course writer or whoever it is who receives the report. There would be little point in providing regular feedback for an organisation that couldn't respond to it. If the student adviser, for example, was told that a part of a course was causing difficulty, what could he do? He might ask the course writer to produce extra material on that topic and send it to the students, or he might ask for a complete section of the course to be rewritten, or he might write a note to the tutors, or he might include something on this problem in a radio programme. If you can list the possible ways in which your organisation might respond to feedback, you can see what sort of feedback would be most useful.

The details of a record-keeping system will obviously depend on the type of activity and the particular arrangements of each organisation. But to illustrate in some detail how a system might work, I will describe how an organisation like LDTC might arrange to keep records on its correspondence teaching. I assume that the organisation has to keep records of it's dealings with students and tutors for administrative purposes. It asks the students to complete an enrolment form before they begin their course; thereafter it keeps some record of the students' progress through the course in order to despatch further materials to the students at the right time and so on. I will show how you might design the forms and the system of record keeping in such a way that information about all the students can be collected at regular intervals.

Who are the students?

If you commission someone to write a new correspondence course, or to write a set of radio programmes to accompany a course, it is helpful to give him some idea of who your students are - of their ages, their previous levels of education, their occupations and so on. You can get this information by putting certain questions on the enrolment form and transferring it to record cards.

The illustration on the next page shows part of an enrolment form based on the one used by LDTC, and the illustration on page 190 shows a student's record card. (At the moment, I am only concerned with the lines of holes along the right hand side and the bottom side of the record card.) When a student is enrolled, his name and address are put on a record card, and certain items of information about him are transferred from his enrolment form onto the record card, using the edge-clipping system (Chapter 7).

The holes along the bottom of the record card are used to record the student's age group (at the date of enrolment), sex, occupation, educational level and the number of years since he finished full-time education. The holes down the side are used to record the subjects the student is taking. If hole A is clipped, it means that the student is taking that subject; if hole B is also clipped, it means that the student began that subject but that he is no longer taking it, either because he has completed the course or because he has 'lapsed'. (A student is described as 'lapsed' i.e. no longer taking the course, if he has not sent in a worksheet in the last eight months.) If a student is no longer taking any of his courses, his card is removed from the pack and placed in the 'old students' file. The

Enrolment form for Junior Certificate (JC) courses

Please complete this form and send it to us. Please write clearly.

Surname (Mr/Mrs/Miss/Ms) ....

Other names ....

Your address ....

....

Date of birth Date ....../Month ....../Year 19....

Your occupation ....

In which year did you complete your full-time schooling? 19 ....

Which standard or form were you in when you left school?

Standard .... Form ....

If you have taken the JC examination, please tell us what result you got:

....

For which subjects do you want to enrol?

(1) .... (2) .... (3) ....

cards in the pack, therefore, belong to students who have sent in at least one worksheet, in any subject, in the last eight months; these are called 'current students'.

Once a year, the records clerk takes the whole pile of current student cards and, with the aid of a knitting needle, sorts them and counts them. He thus produces a statistical profile of the students, in the form shown in the illustration on page 191. If required, he can produce this separately for each subject.

I am not suggesting that every correspondence college should adopt a record keeping system similar to this one. Colleges which offered many more than ten subjects, for instance, would obviously need a different system. A college might also want to record many other things on the student record card, such as students' attendance at one-day seminars. I am only illustrating the point that records can be designed in such a way that statistical information about all the students can be collected at regular intervals without too much difficulty.

STUDENT RECORD CARD

NAME .... ADDRESS ....

....

Subject Date enrolled					Subject Date enrolled					Subject Date enrolled				
Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status

Agric science ○ A ○ B

Biology ○ A ○ B

Book keeping ○ A ○ B

Develop. studies ○ A ○ B

English ○ A ○ B

Geography ○ A ○ B

History ○ A ○ B

Maths ○ A ○ B

Science ○ A ○ B

Sesotho ○ A ○ B

Age group ○ ○ ○

Sex ○

Occupation ○ ○ ○ ○

Age left sch ○ ○ ○

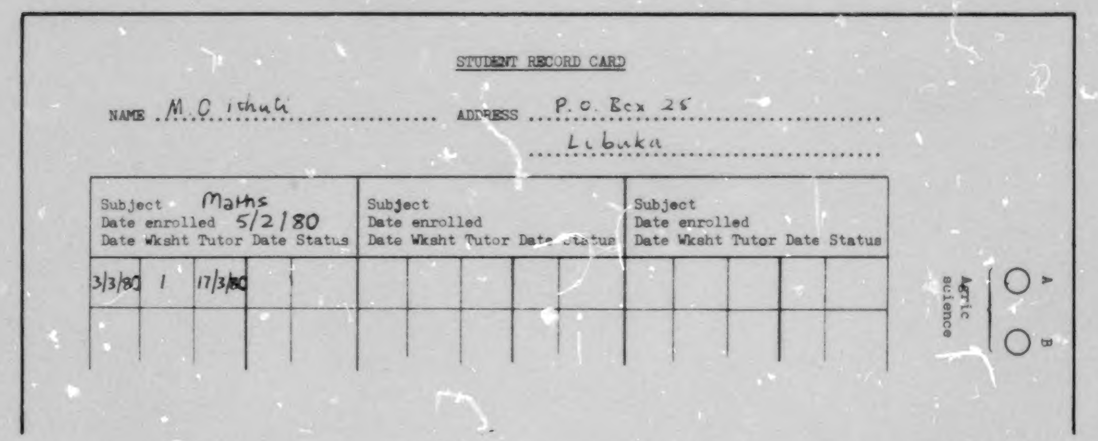
Ed. level ○ ○ ○ ○

Profile of current students. Date ....

		Number	%
Sex:	M		
	F		
			100%
Age group at enrolment:	17 or under		
	18 - 20		
	21 - 25		
	26 - 30		
	31 - 35		
	36 or over		
			100%
Occupation:	Teacher		
	Police		
	Civil servant		
	Shopkeeper/sales		
	Farmer		
	Miner		
	No occupation/unemployed		
	Other		
			100%
Years since left school:	Less than one year		
	1 but less than 3 years		
	3 but less than 5 years		
	5 but less than 8 years		
	8 years or more		
			100%
Previous educational level:	Std 7 or below		
	Form A		
	B		
	C		
	Form D or above		
			100%
Total students			(100%)

How are the students progressing?

As a student works through a correspondence course, he completes a worksheet at the end of each section, which he sends to the college. The college has it marked by the student's tutor and sends it back to the student. This is recorded on the central part of the student's record card. Take the following entry:



STUDENT RECORD CARD

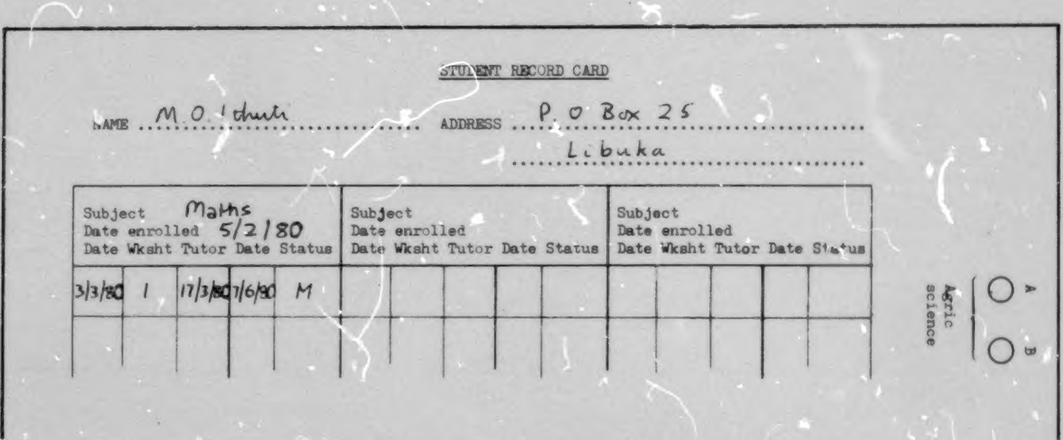
NAME M.C ithuti ADDRESS P.O. Box 25
Libuka

Subject Maths Date enrolled 5/2/80					Subject Date enrolled					Subject Date enrolled				
Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status
3/3/80	1	17/3/80												

Agric science ○ A ○ B

This would mean that on 3 March 1980, the college received Maths worksheet 1 from the student. The worksheet was then sent to the tutor who marked it and returned it to the college. The college received it from the tutor on 17 March 1980 and returned it to the student. (If you wanted, you could include another column on the record card so as to record the student's mark.)

About every six months, the records clerk goes through all the cards and marks each student's 'status'. This 'status' depends on the student's activity over the preceding months. If the student has submitted a worksheet in that subject in the last two months, he is marked A (active). If he last submitted one 2 to 4 months ago, he is marked M (moderately active). If he last submitted one 4 to 8 months ago, he is marked D (dormant). If he has not submitted one in the last eight months, he is marked L (lapsed). Those who have completed the course are marked C (completed). Take the following entry:



STUDENT RECORD CARD

NAME M.O 'thuti ADDRESS P. O Box 25
Libuka

Subject Maths Date enrolled 5/2/80					Subject Date enrolled					Subject Date enrolled				
Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status	Date	Wksht	Tutor	Date	Status
3/3/80	1	17/3/80	7/6/80	M										

Agric science ○ A ○ B

This would mean that the records clerk went through the cards on 7 June 1980; this student had last submitted a worksheet on 3 March 1980, i.e. more than two months ago but less than four months ago, so he was classed as M (moderately active).

For students who have enrolled but not yet sent in their first worksheet, the date of enrolment is used instead of the date of the last worksheet. They are also marked O. For example, a student who enrolled in March 1980 but who had not sent in a worksheet by June 1980 would be classed as MO.

After he has classified all the students in this way, the records clerk counts the number of students with each status for each subject and produces a report on the students' progress, as on the next page.

The form for the report on student progress looks complicated at first sight, but if you read through it carefully you should be able to work out how the different parts relate to each other. The small formula in section 1 for 'number of current students now' simply says that, since the date of the last report, the number of students has been increased by some new enrolments (d), but that you also have to deduct those who have lapsed (b) and those who have completed the course (c). The number of students entered in section 3 as 'current students, not yet sent in any worksheet' should be the same as the number in section 2 that is marked 'f'.

The student adviser can use these reports in various ways. He can check whether student activity is, in a general sense, satisfactory. Let's suppose he had decided that a dropout rate of more than 50% before the first worksheet was unacceptable. If a report showed that more than 50% of current students had not sent in their first worksheet (this would appear in section 2 of the report as f expressed as a percentage of e), he would know that something was seriously wrong and would have to try to find out why. He can compare one subject with another. If activity was markedly lower in one particular subject, he should try to find out what was wrong. As the years go by, he can compare the most recent report with earlier reports for the same subject. This would show if the students' activity was falling off and, again, this would be a warning signal to start investigations.

The reports on student progress give a picture of the recent activity of the students on each course and of their present positions in the course (i.e. the number of worksheets they have sent in). You might also want to know how fast the students are working. One way to summarise this is to calculate the average time between worksheets. Excluding those students who have not sent in any worksheets at all, you note the length of time that each student had been enrolled when he sent in his most recent worksheet, and the number of worksheets that each student has sent in. You then add up the total time since enrolment for all the students and the total number of worksheets submitted by those students. You divide the first by the second to get the mean (average) time between worksheets. To give a simplified example, say Mary had been enrolled for nine months when she submitted her most recent worksheet, which was her fourth worksheet, while John had been enrolled for six months when he submitted his most recent worksheet, which was his first worksheet. In total, they submitted five worksheets in fifteen months, so the mean time between worksheets is 15 ÷ 5 = 3 months.

This is a convenient measure to use for some research purposes. For example, if you were broadcasting radio programmes to your correspondence students, you might want to know whether those students who listened to the

Date .... Report on student progress in subject ....

1. Enrolling, lapsing and completing

Date of last report on student progress ....

(a) Number of current students at that date ....
(b) Students who have lapsed since then .... as % of a ......
(c) Students who have completed since then .... " " ......
(d) Students who have enrolled since then ....
(e) Number of current students now (a + d) - (b + c) ....

2. Current student status

Number of current students (= e, above) ....
Number of active (marked 'A' or 'AO') .... as % of e ......
Number moderately active (marked 'M' or 'MO') .... " " ......
Number dormant (marked 'D' or 'DO') .... " " ....

(f) Have not sent in a worksheet since enrolling ('O' students) .... as % of e ......
(g) Have sent in at least one worksheet (e - f) ....

Of those who have sent in at least one worksheet (= g, above):

Number active (marked 'A') .... as % of g ......
Number moderately active (marked 'M') .... " " ......
Number dormant (marked 'D') .... " " ......

Of those who have not sent in a worksheet (= f, above):

Number active (marked 'AO') .... as % of f ......
Number moderately active (marked 'MO') ...... " " ......
Number dormant (marked 'DO') .... " " ......

3. Worksheet completion

	Not yet sent in any	Last worksheet sent in was number										Total students
		1	2	3	4	5	6	7	8	9	10	
Current (total = e, above)												
as %												100%
Lapsed (total = b, above)												
as %												100%

programmes worked faster than those who didn't. If you could find out which students listened to the programmes and which did not, you could calculate the mean time between worksheets for the radio-listeners and then for the non-listeners and you could see if there was any difference.

I mentioned in Chapter 11 that the mean can sometimes give a misleading impression, and this is true of the mean time between worksheets. It often happens that there are a few students who work very fast while the majority work slowly. To give another simplified example, suppose you had the following data on five students:

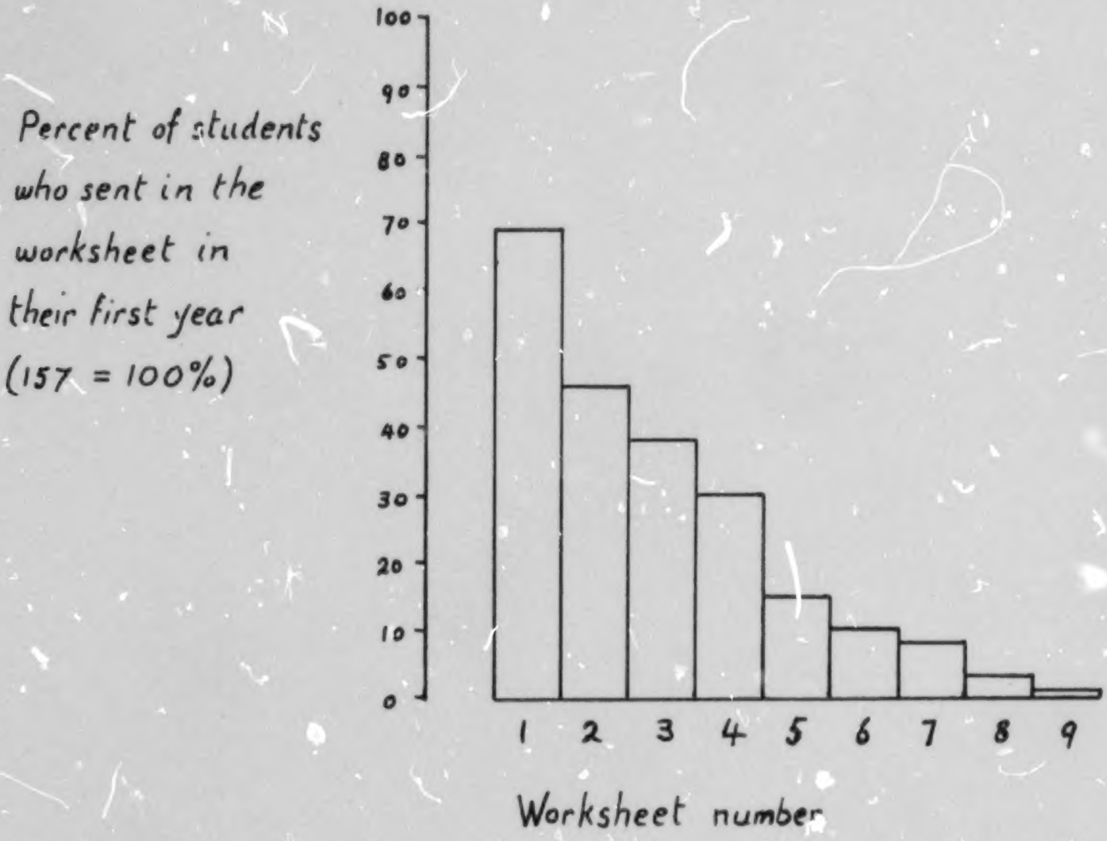
Student A	7 worksheets.	Most recent worksheet 8 months after enrolment.
Student B	2 worksheets.	Most recent worksheet 7 months after enrolment.
Student C	1 worksheet.	Most recent worksheet 6 months after enrolment.
Student D	1 worksheet.	Most recent worksheet 5 months after enrolment.
Student E.	1 worksheet.	Most recent worksheet 4 months after enrolment.
Total	12 worksheets submitted in	30 months

The mean time between worksheets would be $30 \div 12 = 2\frac{1}{2}$ months. But you would be misled if you formed the impression that the typical student was sending in a worksheet regularly every $2\frac{1}{2}$ months. In fact none of the students was doing this; they were sending in worksheets either much faster or much slower than this.

Another way to illustrate the rate of progress is to show how far the students have progressed by a certain time after enrolment - say after one year. To do this, you would take the record cards of those students who have been enrolled for more than one year ('old students' as well as 'current students'). For each one, you note the most recent worksheet he had sent in by a date one year after he had enrolled. For instance, if Mary enrolled on 1 June 1980 and had sent in worksheet number 5 by 1 June 1981, she would go into the group who had submitted 5 worksheets in their first year. You can present this as percentages or as a bar chart, as follows:

Student progress in the first year after enrolment. Maths course.

Out of 157	students who enrolled		(100%)
108	sent in worksheet number	1	(69%)
73	also sent in worksheet number	2	(46%)
60	" " " " "	3	(38%)
47	" " " " "	4	(30%)
24	" " " " "	5	(15%)
15	" " " " "	6	(10%)
12	" " " " "	7	(8%)
5	" " " " "	8	(3%)
2	" " " " "	9	(1%)



You could present these figures in a different way by saying that 49 students (31%) did not send in any worksheets in their first year, 35 students (22%) sent in just one worksheet, 13 (8%) sent in just two worksheets, and so on.

Which parts of the courses need improving?

Correspondence courses do not last for ever; or rather, they should not be allowed to. After a time, a course has to be modified or even completely rewritten, to take account of developments in the subject or changes in teaching method or syllabus or examinations. It is useful to prepare for that event by gathering detailed information about the strengths and weaknesses of the existing course.

One way to get feedback from tutors is to send course monitoring forms to three or four tutors for each course. A tutor fills in one of these forms for each section of the course. Let's say a tutor is filling in a form for section 4 of the maths course. He takes the first ten students who send in worksheet 4 and records their marks on the form for each question of the worksheet. The next page shows one of these forms and also how it might look after it has been filled in. Maximum possible marks per question for this worksheet varied from 2 to 6. Clearly, these students found the worksheet rather difficult, especially questions 4, 6 and 8 ('X' means that the student did not attempt the question). On the back of the form, the tutor writes his own

COURSE MONITORING FORM

Subject .... Worksheet number .... Tutor ....

Question number	1	2	3	4	5	6	7	8	9	10	Total
Max. poss. marks per question											
Student 1											
2											
3											
4											
5											
6											
7											
8											
9											
10											
(Office use) Mean mark per q.											

COURSE MONITORING FORM

Subject Maths Worksheet number 4 Tutor L. I. Palo

Question number	1	2	3	4	5	6	7	8	9	10	Total
Max. poss. marks per question	2	4	2	4	4	4	6	4	5	5	40
Student 1	2	2	1	0	0	X	2	X	1	X	8
2	2	4	2	3	4	4	6	2	4	4	35
3	2	4	1	0	0	X	2	X	1	2	12
4	2	2	0	0	X	X	2	X	1	X	7
5	1	2	0	0	0	2	2	0	0	1	8
6	2	4	2	3	4	3	6	2	5	4	35
7	1	4	2	0	0	X	0	0	1	X	8
8	2	2	1	X	4	X	2	X	2	2	15
9	2	4	2	2	2	X	4	X	3	4	23
10	1	2	0	0	2	X	2	0	2	2	11
(Office use) Mean mark per q.	1·7	3·0	1·1	0·8	1·6	0·9	2·8	0·4	2·0	1·9	16·2

comments about section 4 of the course and about worksheet 4 itself. Then he sends the form back to the college.

The college could also collect feedback from the students by sending a short questionnaire to some of them on each section of the course; the questionnaire might use the techniques I described in the last chapter. A further possibility is to include a single sheet in each coursebook asking students to write any comments. Few get sent back but those few can be useful, since the students who are keen enough to complete them are often lively and critical.

Thus, as time goes by, the college accumulates a lot of detailed information about every section of every course. Looking at the forms, you can pick out those questions which are giving trouble to the students. The person who is modifying the course can look at those questions and try to see whether the question itself is too hard or whether the lesson failed to give the student enough help to answer the question. In addition, he has the comments of three or four tutors and a number of students on every section of the course.

Exam results

Many correspondence courses are intended to prepare students for examinations. LDTC, for example, offers courses up to the Junior Certificate level (the exam that school students take after three years of secondary school) and to the General Certificate of Education 'Ordinary level' (the exam that school students take after five years of secondary school). The success of these courses will be judged, to some extent, by the students' exam results.

Before discussing the difficulties this raises, I should point out that examination results are not the only criterion of success. Every teacher knows that there is a difference between teaching and 'cramming'. Good teaching - and this applies to correspondence tuition just as much as to classroom teaching - is more than just preparing students for exams. So a college that gets its students through exams is not necessarily doing a good job in the sense of providing good tuition. On the other hand, a college that got none of its students through exams would probably not be doing a good job in any sense at all. The students, certainly, think that exam results are important. So, though exam results are not the only criterion of success, they are an important one.

The first problem in measuring exam success might be simply finding out what results the students get. If the examining body is separate from the distance-teaching institution, the students can probably apply to take the exam, as private candidates,* at any time they choose. The college will not necessarily know that they are taking the exam at all. One solution might be for the examination board to put a question on the application form for the exam, 'Are you enrolled as a student with the following college(s)... ?'

---

* 'Private' candidates are those who make their own application to the examining board to take an exam, rather than being entered for the exam by a school or college.

Then, perhaps, the board could produce a separate list of results for students of each college. Failing that, you can encourage your students to tell you when they are taking exams by offering some extra service - a booklet or a one-day course or something - which is specially for students about to take exams. Whether you can find out the students' results after the exam will depend on how exactly the results are published.

Assuming that you can find out your students' exam results, the next problem is to convert these into some sort of pass rate. This is more difficult with correspondence students than it is with school students. Suppose that a school has sixty students in their third year; all sixty take an exam and thirty pass. The pass rate is 50% - thirty out of sixty. But suppose that sixty students from a correspondence college take an exam and thirty of them pass. It could be very misleading to say that the pass rate is 50% - thirty out of sixty. Closer examination of the figures might reveal that only six of the sixty had actually completed the course, and all six passed. Eighteen had completed most of the course (but not all of it) and twelve of them passed. Twenty-four had only done a little of the course, and nine of these passed, whilst the remaining twelve had not yet sent in any worksheets at all, and three of these passed. You might argue that the real pass rate was 100%(of the six who had completed the course, all passed), or that it was 75% (of the 24 students who had completed all or most of the course, 18 passed).

But there is still a further complication. Suppose that a lot of the students who had enrolled in the subject did not take the exam. There were twelve students who had completed most of the course but who did not take the exam; there were another 60 who had done a little of the course, and there were another 78 who had not sent in any worksheets since they enrolled. Perhaps the pass rate should be 50% (of the 36 students who had completed all or most of the course, 18 passed the exam), or perhaps it should be 14% (of the 210 students who had enrolled in this subject, 30 passed the exam).

So the pass rate, in this example, could be 14%, 50%, 75% or 100%, depending on how you calculated it. If a correspondence college claimed that it had a pass rate of 80% (or whatever %), you should ask them how exactly they arrived at this figure. The fairest thing to do would be to present all the information I have given in the form of a table, something like this:

	Total	Took the exam	Passed the exam
Had completed the course	6	6	6
Had done two-thirds of the course but not all of it	30	18	12
Had done some of the course but less than two-thirds	84	24	9
Had not sent in any worksheets since enrolling	90	12	3
Total	210	60	30

If you wanted to present any of these figures as percentages, you should make clear, as always, what the base totals were, e.g. 'Out of 36 students who had completed all or most of the course (two-thirds of the course or more), 24 (67%) took the exam; of these 24 students, 18 (75%) passed it.' (Percentages are discussed in Chapter 11.)

Having arranged the figures so as to give a full and fair picture of the students' exam results, the next problem is to decide whether the results are satisfactory or not. This means comparing them with something. Is a pass rate of 50% good or bad? If you were expecting 30%, then it's good; if you thought it should have been 80%, then it's bad. There are various figures that you might compare your results with.

It is tempting to compare them with the results obtained by school students. Someone might say, 'Of all the school students who took this exam, 40% passed; of our correspondence students who took this exam, 60% passed, so we have done better than the schools.' But such claims do not stand up to close inspection. The 'pass rates', as I have just explained, are not comparable. In Lesotho, for example, only a minority of secondary school students drop out before completing the three-year course to the Junior Certificate level whereas a high proportion of LDTC students either drop out and fail to take the exam or take the exam before completing the course. The two systems are different in other ways. Schoolchildren have to pass a qualifying examination before entering secondary school, whereas anyone can enrol for a correspondence course with LDTC. Secondary school students are teenagers; most of LDTC's students are adults. So, if you found a difference between the 'pass rate' of school students and the 'pass rate' of correspondence students, it would prove merely that the students were different sorts of people or that the rates were calculated in different ways. It would not prove that the correspondence college was doing well, or badly.

It might be more useful to compare the results of your correspondence students with those of private candidates who had received other sorts of assistance, such as evening classes; or you might compare them with candidates who had received no assistance at all. You would probably be comparing groups of students who at least resembled each other in age and so on. Even here, however, it would be difficult to decide which of the possible pass rates was the appropriate one to use.

Most useful would be to compare the results of your students with those of students from other correspondence colleges, being careful to calculate the pass rates in the same way for each college. Finally, you could at least compare one year's results with those of previous years. This would not tell you whether you were doing well or badly by the standards of correspondence education in general, but it would tell you if you seemed to be doing better or worse than you had done in the past.

In this section I have strayed into the topic of the next chapter; when you use exam results to measure the success of a correspondence course, you are doing evaluation. It is a bit different from the kind of evaluation normally undertaken by a distance-teaching institution, in that the testing of the students is designed and carried out by an independent body (the examination board). Nonetheless, it has raised two questions which are central to evaluation: 'To what extent are the exam results an appropriate measure of what we were trying to achieve?' and 'What shall we compare our results with, in order to decide whether they are satisfactory or not?' I will take them up again in Chapters 15 and 16.

15 Evaluation: looking at projects as experiments

This chapter considers one aspect of evaluation - finding out how far a project is achieving what it is supposed to achieve.

Evaluating the education, not the students It is not the students, as individuals, who are being graded or evaluated, but the education they have received.

What is the desired effect? You need to clarify what the project is supposed to be achieving before you can assess whether it is achieving it.

Did the programme run as it was meant to? Part of the evaluation consists of finding out simply whether the work was done as it was supposed to be.

Are the learners identified in advance? The way you do an evaluation depends partly on whether you know who the students are at the outset.

How many people have we reached? A survey, or an appeal to the students, can give you an idea of how large your audience is.

Did the project have any effect at all? You can't assume that, if your students know something, they've learnt it from your course. You need a point of comparison.

Before and after A baseline survey conducted before a campaign provides a point of comparison for survey results after the campaign.

Control groups It may be possible to compare the people who have received the instruction with an equivalent group who haven't.

Identified audiences The before-and-after design and the control-group design can be adapted for use with audiences that are identified before the instruction begins.

A weak design is better than nothing Evaluations that use weak designs - as they often have to - can still be useful.

What to measure You measure the effects of a project by observing things or asking questions or giving people tests. You have to get as close as possible to the important effect.

Was the effect big enough? Statistical significance is not the sole criterion. The effect must be judged against the project's objectives.

In action research, you conduct an action project as an experiment. You have an idea for a scheme - for example that booklets would be an effective way of getting practical instruction to rural people - and you test the idea by actually putting it into practice (i.e. publishing some booklets) and seeing if it works. You know what you want to achieve in your action project; the research consists of finding out to what extent you actually do achieve what you want to achieve.

One approach to evaluation is to regard all projects as if they were experiments. You look at the project and you say, 'This is what it was supposed to achieve. Did it work?' Another way to describe this approach to evaluation would be to say that you are judging the project on its own terms; you are measuring how the project matches up to its own aims.

Evaluating the education, not the students

The terms 'evaluation', 'measurement' and 'assessment' are often used, in education, to mean 'grading the students'. Schools are constantly ranking their pupils in order of achievement. Many schools give tests at the end of every term to decide who is top of the class, who is second, who is third and so on. Exam results, such as those for the General Certificate of Education, are often given as grades - those who have done best are given Grade A, the second best Grade B, and so on down to those who have done worst, who are given Grade F. Clearly, in this sort of evaluation, it is the students who are being evaluated.

Increasingly, however, the term 'evaluation' is being used to describe some assessment of the education that the students have received, not of the students themselves. Suppose a teacher thinks that his students are being held back by their slow reading speed, so he decides to give them a course in rapid reading. At the end of the course he tests their reading speed to find out whether it is any faster than it was before. Now, although he is giving the students a test, the purpose of the test is not to grade the students in order of their reading speed, but to assess whether the course has been effective. It is the course that he is evaluating, not the students. This is the kind of evaluation that I am talking about in this book - evaluating the education, not grading the students.

What is the desired effect?

A distance-teaching project, obviously, is undertaken with some purpose, perhaps many purposes, in mind. It may be that the people undertaking it have never spelt out clearly what the purpose is, but, nonetheless, they intend that the project should have some effect on people. They hope that, as a result of the project, people will know more about some subject, or feel differently or take some action, or something. The approach to evaluation that I am covering in this chapter can be summed up as an attempt to answer the question 'Is the project having the desired effect?'

Sometimes, as in the case of the rapid reading course, it is easy to say what the purpose of the project is. But often it is not so clear. If you are offering a correspondence course in physical science, do you just want to help students to pass the exam, or are you trying also to arouse their interest in science or to encourage them to adopt a scientific attitude towards problems? If you are conducting a health education campaign on malaria, are you trying to teach people certain facts, so that they could give the right answers to questions about mosquitoes and so on, or are you trying to change their traditional beliefs about how diseases are caused, or are you hoping that they will take certain steps to combat malaria, such as draining swamps where mosquitoes breed?

Clearly, in order to answer the question, 'Is our programme having the desired effect?' you need to have a fairly precise idea of what you mean by 'the desired effect'; you have to be able to say what it is that you are trying to achieve. You may find that people are not clear about what the project's aims are, or even that they disagree about the aims. A useful by-product of evaluation is to bring these matters into the open.

In an earlier section ('Clarifying objectives', in Chapter 12) I described a way of approaching this question. At that point I was talking about particular instructional materials, such as a single correspondence lesson or a pamphlet or a radio programme, but the same principles can be applied to a whole project. Without describing the details again, the main idea is that you try to specify what you want the learners to do, or to be able to do, as a result of the project. The important word here is 'do'. Of course, your project might not be directed primarily at getting the learners to do anything. You might be trying to increase their knowledge, arouse their interest, or change their beliefs or attitudes. In other words, you are mainly concerned about what goes on in their minds. But you cannot see what someone knows or thinks or feels just by looking at him; you cannot get inside his head to measure what effect you've had. You can only see it if you get him to do something - to answer a question or solve a problem or perform some task. The evaluator's question 'How will we measure the success of our efforts?' leads naturally to the question, 'What exactly are we trying to achieve?' If the educator answers that question by giving examples of what he wants the students to do, or to be able to do, as aresult of the project, he is also providing the evaluator with something he can measure. (When I use the terms 'researcher' and 'evaluator', I am talking about the same person. The researcher is sometimes called 'the evaluator' if he's doing evaluation.)

Did the programme run as it was meant to?

When planning how you will measure the effect of an educational programme, you are inclined to assume that the programme will operate in the way it is meant to, but this assumption can be wrong. Programmes often fail to operate in the way they are meant to. The fault can lie with your own organisation; perhaps parts of a correspondence course don't get printed in time for when the students need them, or perhaps a radio programme does not get recorded in time for broadcasting. But the fault is most likely to occur if you depend, for part of the project, on some other people who are less well informed about it or less committed to it. If you deliver your tapes to the radio station, the tapes might not get broadcast. If you circulate parcels of booklets for fieldworkers to sell, the parcels might never get opened.

Recognising this simple point can save you a lot of trouble. An example is a piece of work we did at LDTC for an agricultural project; we edited and printed a series of monthly newsletters, six in all, for farmers in the project area. We designed a questionnaire to assess the effects of these newsletters on the farmers and we interviewed a number of the farmers. All we discovered was that the agricultural project had failed to distribute the newsletters.

So you should begin your evaluation by finding out simply whether the project has taken place in the way it was intended to. Were the radio programmes actually broadcast? Were the leaflets distributed? Were the posters displayed? If you find out that the materials did not even reach the audience, you do not need to go on.

I emphasise that such questions are only the beginning of an evaluation. Many organisations seem to think that, once these questions have been settled, the evaluation is complete. One reads reports such as, '5 000 leaflets were distributed,' or 'Sixty people enrolled for the bookkeeping course,' as if that is all that needs to be said. But the important questions still remain - 'What did the people learn from the leaflets?' 'Were the participants any better at bookkeeping after the course than they were before?'

Another commonsense point to consider early in the evaluation is whether it is even reasonable to hope that the programme might have had the desired effect. Projects that begin with great ambitions can fall short in the execution. Perhaps a series of twelve booklets were to be published, with 100 000 copies of each, but, in the event, only two were printed, at 5 000 each. Perhaps the course was intended to reach 250 volunteer health workers, but actually reached only 17. In such a case, it would be foolish to conduct elaborate field research to demonstrate that the programme has fallen short of its aims. That is already obvious. Rather you should try to find out why the programme fell so far short of its aims, and then you should reduce the scale of the evaluation to make it more appropriate to the actual size of the programme. For example, you might still want to find out how much impact the 5 000 booklets had, or how much effect the course had on the 17 volunteers who actually took it.

Are the learners identified in advance?

In my earlier example of the teacher giving his students a rapid-reading course, the teacher obviously knows in advance who the students are. And this is true of some kinds of distance teaching. When students enrol for a correspondence course, for example, you know who the students are before they start work on the course; you can give a list of names of the people whom you are trying to teach. But this is not true of some other kinds of distance teaching. Suppose you broadcast a radio programme about vaccinating babies. You know the sort of people you want to reach - parents of young children. But you do not know in advance who is going to listen to the programme. You cannot give a list of the names of your 'students'; you do not know who they are going to be.

This difference affects the way you do an evaluation; it affects both the questions that you ask about the project and the research designs that you can use. First I'll deal with cases like the radio programme on vaccination, where you do not know in advance who the learners will be.

How many people have we reached?

If you broadcast a radio programme, you want to know how many people listen. If you publish a leaflet, you want to know how many people read it.

If you are expecting your material to reach more than about 5% of the total possible audience, you can find out by doing a sample survey (see Chapters 4-9). To assess the audience for, say, a single radio programme about vaccinating babies, you would have to do your survey in the few days following the broadcast. You would locate a sample of the sort of people you were trying to reach, such as parents of young children, and find out how many of them had heard the programme. As I mentioned in the chapter on questionnaires (Chapter 5), it is not a good idea to ask 'Did you listen to a programme about vaccination on Monday evening?' People will realise that the interviewer is connected with the people who broadcast the programme and will tend to say Yes out of politeness. It would be better to ask:

> Did you listen to the radio on Monday evening?
>
> (If yes) Which of these programmes did you listen to -
> News at Six?
> Farming Report?
> Family health?
> The Monday play?

You can also check on whether people who say they have listened to a programme really did so by asking them some questions about it. These should be questions which people can easily answer if they have listened to the programme, but which they wouldn't be able to answer if they hadn't, such as 'The programme included a short drama; who were the main characters?' Similarly, if you published a leaflet for farmers, you could find out if farmers really had seen it by asking such questions as, 'What colour paper was it printed on?' or 'What picture was on the front?'

If you are expecting your material to reach only a tiny proportion of the total possible audience, the sample survey is no good. Suppose you are broadcasting a series of programmes to teach a foreign language and you think that about one person in a thousand is listening. If you did a sample survey, you would have to interview hundreds of people to find just one student. Obviously this would be far too costly.

There is no easy way to measure the size of the audience in cases like this. One possibility would be to ask the students to write to you, and you could encourage them to do so by offering some incentive; for example, you could offer a free pamphlet to accompany the programmes, or you could ask a question in the programme and offer a prize for the first correct answer. The weakness of this, of course, is that not all the students will write in. In order to make a guess about the number of students on the basis of the number of letters, you'd have to have some idea of the proportion who write in. But at least it would give you a lower figure; you could say 'We know that at least 200 people (or whatever) are following this course.'

Did the project have any effect at all?

As a part of an evaluation of the impact of a cookery book that we published at LDTC, we interviewed some housewives who had had the book for a few months and who said they had read it. We asked them some questions, among other things, about the nutritional value of beans. Nearly all of them gave the right answers. This seemed most encouraging. Clearly, the book had taught them something. Or had it? We also interviewed some housewives who had not read the book; nearly all of these housewives also gave the right answers. The book had not taught anyone about the nutritional value of beans; most housewives knew all about this anyway.

This example illustrates the serious weakness of one research design which is often used in educational evaluation. If the researcher confines his attention to the students, and gives them a test only after they have received the instruction, he has nothing with which to compare their results. Suppose we had interviewed only housewives who had read the book and we had calculated their score on the test, e.g. '93% of these housewives answered correctly the questions about beans.' How could we have interpreted this result? We would have been tempted simply to assume that, without the booklet, they would not have answered so well. And we would have been completely wrong.

If you give people a test after your project has taken place, these results on their own do not tell you whether your project has had any effect. You need some other figures to compare them with. There are, basically, two different comparisons you can make. One is to give people a test before the project as well as after it, e.g. 'A sample of people who were interviewed before our campaign about TB got a mean (average) score of two out of ten in a test of knowledge about this disease; a sample of people who were interviewed after the campaign got a mean score of five out of ten.' The other is to give the test to some people who have not received the instruction as well as to those who have, e.g. ' People who had listened to the radio programme on prevention of TB got a mean score of four out of ten on a test of knowledge about this disease, whereas people who had not heard the programme got a mean score of one out of ten.' The next two sections consider each of these methods in turn.

Before and after

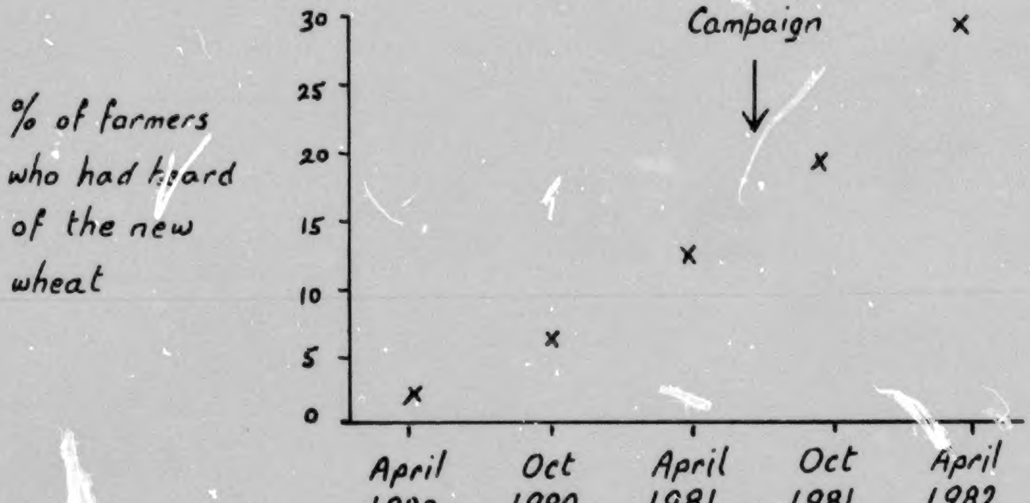
If the distance-teaching programme is intended to reach a lot of people, you can conduct a sample survey before the programme begins, and then repeat the survey after the programme has been running for some time, or after it has ended. For example, if you were trying to inform farmers about a new high-yielding type of wheat, and you hoped to reach at least 10% of the farmers in the country, you could interview a sample of farmers before the campaign, to find out how many of them knew about this type of wheat, and then interview another sample after the campaign, asking them the same questions. The 'before' survey is often referred to as a 'baseline' survey, which indicates that its purpose is to provide a point of comparison against which to judge the results of the 'after' survey.

When using this research design, it is important to do the two surveys in the same way, drawing the sample in the same way, asking the questions and recording the answers in the same way, and so on. The reason is obvious. If you took a 'before' sample of illiterate, isolated, traditional farmers and an 'after' sample of literate, radio-owning, progressive farmers, you would find a difference in the proportion who had heard of the new wheat, but this would not necessarily be due to your campaign.

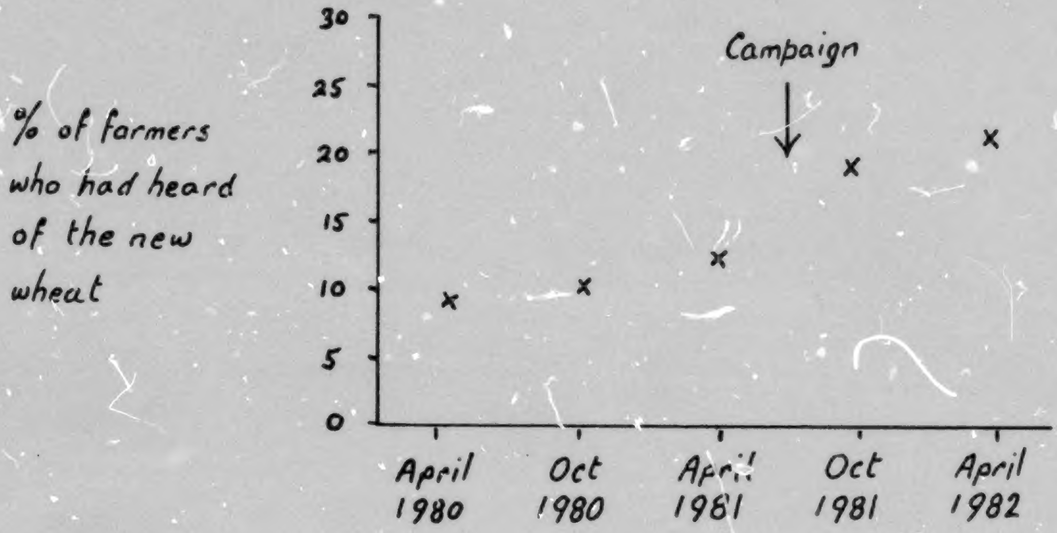
The main weakness in this research design, even when the two surveys are carried out in the same way, is that, if you find a difference, you cannot be sure that your campaign can take the credit. Let's say that the two surveys of farmers were separated by six months. Many other things, in addition to your campaign, will have happened during those six months. Say that 12% of the farmers in the baseline survey had heard of this new wheat, and 18% in the 'after' survey. It looks as though your campaign has had an effect but you cannot be certain that it was, in fact, your campaign that made the difference; the farmers might have heard about the wheat in some other way.

In the chapter on experiments (Chapter 10), I explained the concept of internal validity. The type of evaluation I am describing involves treating an action-project as if it was an experiment. Hence, the problems that arise in drawing conclusions from experiments also arise in drawing conclusions from this type of evaluation. To continue the example of the wheat campaign, you have your explanation of the results ('The increase from 12% to 18% shows that our campaign had an effect'), but it is not the only possible explanation of the results, and it is not necessarily the right one. Someone else might say 'It had nothing to do with your campaign; the farmers were simply spreading the word among themselves,' or, 'The seed salesmen travelled all over the country giving away free samples of the new seed; the farmers came to hear of it from the salesmen, not from your campaign.' The before-and-after evaluation is a weak experimental design in that the results are wide open to rival interpretations. You might consider, for other reasons, that the rival explanations are unlikely to be right, but the research design itself does not exclude them.

One way to improve on this design is to increase the number of times you do your survey. Suppose you interviewed samples of farmers in three separate surveys before the campaign and in two surveys after the campaign. If you put the results in the form of a graph, they might look like this:



Bad news for the campaign organisers. Knowledge of the new wheat was increasing steadily anyway; the campaign had no effect at all. But suppose the results looked like this:



This is more encouraging. Obviously, something had an effect on the farmers between April and October 1981, and it could well have been the campaign. But note that even the second diagram does not prove that it was the campaign that had the effect. It is still possible that something else happened at the same time as the campaign, such as the handing out of free samples of seed, and that it was this other thing, rather than the campaign, that affected the farmers.

When the same statistic (in this case 'Percentage of farmers who had heard of the new wheat') is gathered at different points in time, the resulting set of figures is called a 'time series'. The main problem with doing the same survey several times is that it is hard work. Indeed, a common problem even with the simpler before-and-after design is that the 'after' survey never gets done. People are enthusiastic at the start of a project and do a thorough baseline survey, but, when the time comes to do the 'after' survey, they are tired of it (or perhaps the original people have left the project) and they can't be bothered to do it.

If it is not possible to generate your own time series by doing a repeated survey, you can sometimes make use of time-series figures that have been collected for other purposes. For example, family-planning clinics might keep monthly records of the number of people attending the clinic; if you were evaluating the impact of a family-planning campaign, you might use these figures. Or the cooperative society might keep records of sales of fertiliser; you might use these to evaluate an agricultural programme. While such figures can be very useful, you should approach them with some caution.

First you should try to assess how reliable they are. Who actually collects the figures at village level, and how do they do it? Have they any reason to make the figures higher or lower than they really are? Are the figures collected in the same way from month to month, i.e. can you compare one month's figures with another month's?

Secondly, you should consider whether other factors are affecting the figures in addition to your educational programme. Let's say that clinic attendance goes up in the month after your campaign. Perhaps the family-planning organisation has recently introduced an incentive scheme for its fieldworkers. Or say that fertiliser sales increase sharply after you have distributed a leaflet recommending fertiliser. Perhaps sales always go up at this time of the year. Generally people are quick to seek alternative explanations for disappointing figures (e.g. 'Fertiliser sales actually declined somewhat during the period of the project, but we must bear in mind that the price of fertiliser had gone up'). But they are reluctant to do so when the figures seem to be encouraging ('Our success is indicated by the increase in fertiliser sales during the period of the project'). This is understandable, but it is not very scientific.

Control groups

The second way to find out if a distance-teaching project has had any effect is to compare some people who have received the instruction with some people who haven't. The people who have not received the instruction are known as the 'control group'. I described this research design in Chapter 10. I explained there that, ideally, you want two groups that were equivalent in every way before the instruction, and that the best way to achieve this is to assign people at random to the two groups.

If you are broadcasting radio programmes, distributing leaflets or offering booklets for sale, how can you possibly assign people at random to the experimental group (who will receive the instruction) and the control group (who won't)? Take the booklets example. You would have to wait until people showed an interest in buying the booklet. Then you'd decide for each person, at random, whether to sell him the booklet or not. Then you'd have to prevent the non-buyers from getting hold of the booklet. If you could possibly do all this, then you'd have your two equivalent groups. But, obviously, you couldn't do it.

With this sort of distance-teaching project, you cannot assign individual people to the two groups. But what you can sometimes do - and it's almost as good - is to assign collections of people at random to the two groups. If you were offering booklets for sale through clinics, you could randomly divide the clinics into two groups and offer the booklets in the first group but not in the second. If you were distributing leaflets, you could select villages at random and distribute them in some but not in others. If the country was divided into regions which had different radio stations, you could select regions at random and broadcast your programme in some but not in others.

To continue the example of the wheat campaign, let's suppose you were using mobile film shows backed up by leaflets. You might select, at random, an experimental group of thirty villages and a control group of thirty. (The experimental and control groups do not have to contain exactly the same number of villages, in fact.) The film van visits the thirty experimental villages, showing the film and distributing the leaflets, but takes care not to visit the control villages. Then you conduct a survey in all sixty villages to find out how many farmers have heard of the new wheat. You find, let's say, that 38% of the farmers, on average, in the experimental villages have heard of the new wheat as against 12% in the control villages.

If you were worried about the possibility of farmers in the experimental villages passing on the message to farmers in the control villages (a problem that researchers call 'contamination'), you could decide, when drawing the sample, that no two villages should be less than, say, five kilometres apart. Or, looking at it more positively, you could try to find out, in your survey, to what extent this had happened; the campaign organisers would be interested to know.

People sometimes raise ethical or political objections against this research design. They argue that it is not fair to withhold information from the unfortunate people who happen to be in the control villages; besides, the villagers might object if they found out that they were being discriminated against in this way. But you often get the opportunity to use this research design in the early stages of a campaign, in a way which is not at all objectionable. Suppose the film van is going to visit 400 villages in all, doing about two a day. It can't visit them all in the first two weeks. All you require for your experiment is that the van should visit the thirty experimental villages early in the campaign and leave the control villages till later, leaving you some time to do your survey in between.

Very often, however, the researcher has no control over who receives the instruction (or who receives it first) and who doesn't, especially when radio or TV is being used, so he cannot use the randomised control group design. If you can't arrange an equivalent control group, you have to make do with a non-equivalent one. Perhaps radio reception is bad in one part of the country; a sample of villagers from that part of the country would provide a useful control group against which to compare samples from other parts of the country when evaluating radio. Perhaps a campaign begins in one district with the intention of spreading to others later; you could compare people in the first district with people in districts that have not yet been reached.

If all else fails, you can compare some people who have received the instruction with some who haven't - people who listened to the programme and people who didn't, people who bought the booklet and people who didn't, and so on. The great weakness with this is that the two groups are certainly not equivalent; in fact they are not even near to being equivalent. Suppose you interview some people who listened to your programme on vaccination and some people who didn't. Let's say you find a big difference in their scores on a test of knowledge about vaccination - the listeners scoring eight out of ten, on average, as against the non-listeners' three out of ten. Does this show that your programme had a big effect? Not necessarily. Radio owners, who form the majority of the listeners, are likely to be richer, more educated, more literate and generally better informed than others. People who listened attentively to the programme are more likely to be in favour of vaccination and interested in the topic anyway. So you'd expect your two groups of respondents to get different scores on the test, even if you had never broadcast your programme.

The same problem would arise if you interviewed some people who had bought a booklet and some who hadn't. Perhaps you interview 100 people who have bought the booklet and you find that 90% answer your questions correctly, whereas you interview 100 who did not buy the booklet and you find that only 60% answer the questions correctly. It looks as though the first group learned something from the booklet. But perhaps the first group are better educated - such people are more likely to buy booklets; you might have found this difference in test results between the two groups even if the booklets had never been published.

One way to improve slightly (but only slightly) on this design is to get two groups who are matched in certain respects. When evaluating your radio programme, you might restrict your attention to radio-owners; if someone doesn't own a radio, you don't interview him. And you compare those radio-owners who did listen to your programme with those who didn't.

Or you can get groups which are matched in several respects. Say you are interviewing people about a booklet on vegetable-growing; some have read the booklet and some haven't. In addition to the test questions, you also record the sex, age and educational level of the respondents, and you try to assess their interest in growing vegetables by asking whether they have ever grown vegetables in the past (i.e. before the booklet was published). In the analysis, you take a subgroup of the booklet-buyers, such as women aged 20 to 40 with an education of Standard 5 or higher who have never grown vegetables before. You then take the corresponding subgroup of the non-buyers. These two subgroups are now matched, at least in these characteristics, and you compare their test results (see the section 'Hidden Factors' in Chapter 9.)

It is worthwhile trying to match your groups with respect to one or two factors that you think are important to the topic in question; you are comparing groups which are equivalent at least in those respects on which you've matched them. But you must not make the mistake (a common one) of thinking that, because you've made your groups equivalent in certain respects, you can assume that they are equivalent in all other respects. They aren't. However many ways you make them equivalent (in sex, age group, educational level, radio ownership and so on), there are bound to be many ways in which they are different, and any of these differences could give rise to a difference in their test scores, irrespective of your booklet or radio programme.

Identified audiences

Now I'll deal with those situations in which you know who the learners are before the teaching begins. You might want to find out how much a class of schoolchildren learn from a single radio programme that they listen to in class, or how much the members of a radio learning group learn from one meeting. Or perhaps you want to measure the effect of a film show or a lecture by an extension agent.

As with the unidentified audiences, it is not enough just to give the learners a test at the end and to assume that any knowledge they display must have been acquired from the tuition they've just had. You need a point of comparison.

Again, you can use the before-and-after design. You ask the people some questions before the instruction and then again at the end and you see if there has been any change. This design is appealingly simple but it contains a flaw. People who have taken a pre-test will approach the instruction in a different frame of mind from people who haven't. Suppose the pre-test contains five questions on crop-rotation. When the people listen to the programme (or watch the film or whatever), they will already be thinking of crop-rotation and perhaps looking out for the answers to these questions. Indeed, putting questions to the students at the start of a lesson is a well known teaching device, for precisely this reason.

Take the example of a film about family planning which presents the message indirectly though a drama of family life. Let's suppose that the film-makers have been too indirect and that, in ordinary showings of the film, half the audience go home without it crossing their minds that the film had anything to do with family planning. A researcher who gave everyone a pre-test, with questions about family planning, would never discover this. The people would gather from the pre-test that this was a film about family planning so they would be looking for a family planning theme in the film.

You can sometimes reduce this problem by giving people the pre-test some time in advance of the instruction. With a class of schoolchildren, for example, you could give them the pre-test a week in advance, so that it wouldn't be in their minds when they listened to the programme. But a better solution is to give the pre-test only to some of the audience, selected at random, and the post-test to others. (Or you can give the post-test to everyone and compare the results of those who took the pre-test and those who didn't, to see if the pre-test had any effect. If it had no effect, you can use everyone's post-test results. If it did have an effect, you discard the post-test results of the people who took the pre-test.)

The examples I have given so far in this section have been single sessions of instruction, like one radio programme or one film show. Instruction which extends over weeks or months, such as a radio learning-group campaign or a correspondence course, poses further problems. One is that the educators hope that the students will learn a great deal from the course, or at least that the keen ones will. A test which might be appropriate as a pre-test would be much too easy as a post-test. You can get round this problem by producing two different tests but keeping some items the same in both. The easier items in the pre-test give you an idea of the students' level at the beginning of the course. The harder items in the post-test enable you to see how well the better students have done by the end. And the items which are the same in both (known as 'anchor items') provide you with a point of comparison.

Another problem with extended courses is that not all the students who start the course will complete it. So, in addition to measuring the achievement of the students who stay on the course, you have to make some measurement of the dropout. Suppose you were told that people who had taken a basic literacy course had progressed from being unable to write their own names to being able to write letters and read newspapers. Obviously you also need to know the dropout rate. It makes a difference to your assessment of the course whether the proportion of beginners who reach the end is 70%, 20% or 1%. (For more detail on this, with reference to correspondence students, see Chapter 14.)

With identified audiences, as with unidentified ones, the randomised control group design is the best if you can do it. The same ethical and political problems arise. If you were offering a correspondence course, or a literacy course, it would be unfair to refuse to teach a randomly selected group of would-be students, just because the researcher wanted a control group. But the opportunity to use this design might arise if the number of would-be students exceeded the number of places on the course. Rather than adopt the policy of first-come-first-served, you could receive applications up to a certain date and then select at random the ones to be accepted on the course. Besides being fair to everyone, this would provide the ideal basis for experimental evaluation.

A weak design is better than nothing

Of the research designs I have described, only the randomised control group design is completely satisfactory. If you have two groups that are equivalent in every way except that one receives some instruction and the other does not, and if you find that they perform differently on some test after the instruction, you can be confident that the instruction had an effect. All the other designs have weaknesses. The trouble with a weak design is that the results can be misleading. They might make it look as though the instruction has had an effect when in fact it hasn't. I have described them because you often have to make do with a weak design, and even an evaluation which uses a weak design is better than no evaluation at all.

First, it is always possible that the results will show that the instruction has had no effect. Providing that your measurement device (i.e. your set of test questions or whatever) is reasonably sensitive, this result is convincing even with a very weak design. I gave the example earlier of LDTC's evaluation of a cookery book. We located some women who had read the book and some women who hadn't and we asked them some questions. This was a very weak design; if the booklet readers had answered the questions better, we could not have concluded confidently that it was because of our booklet - these two groups of women were different in many ways. But in fact the booklet readers did not answer the questions better. If the book had taught them anything, they should have answered better. They didn't, so we can conclude with confidence that it didn't teach them anything (at least on the topics tested by these questions).

Secondly, I have emphasised the weaknesses of certain designs in order to discourage people from jumping to conclusions. If you have used a weak design and the results suggest that your distance teaching has had an effect, you should always ask yourself whether there is some other plausible explanation of the results. But I don't want to suggest that you can never draw conclusions confidently from weak designs.

Suppose you have been encouraging farmers to adopt a new type of shelter for storing crops. You have been promoting this type of shelter over the radio and inviting farmers to write in for a booklet on how to build it. A baseline survey showed that no farmers at all were using this type of shelter, and a survey one year later shows that 5% of farmers are using it. You can see that these new shelters are the type you've been promoting. You know that no one else has been promoting them simultaneously. And the farmers with the new shelters tell you that they got the idea from your radio programmes and wrote in for your booklet. Even though you have used the 'before-and-after' design, which, as I've explained, has certain weaknesses, you would not have any doubts about drawing the conclusion that your distance teaching has had an effect. Of course it has.

If an experimental evaluation, using a weak design, has produced results which suggest that your distance teaching has had an effect, you should force yourself to think of other factors that might have produced this effect. Then you weigh your first explanation of the result ('The distance teaching produced the effect') against the other possible explanations of the result. How much faith you have, finally, in your first explanation depends on the strength of the rival explanations.

What to measure?

So far, I have discussed the general design of an experimental evaluation, but I have not said much about how you actually measure whether the distance teaching has had an effect on someone or not. I have talked vaguely about asking people questions or giving them a test. But of course it is important that your measuring device - the questions or the test or whatever - should be appropriate for the job. Suppose an institute runs a very basic course in numeracy for pre-school children, intended simply to introduce the most elementary concepts of numbers and patterns, and the evaluator, completely out of touch with the teachers, tries to measure the effect by giving the children multiplication problems to solve. He finds that none of the children can solve the problems, and pronounces that the course has had no effect. The teachers will complain that the test was inappropriate.

Sometimes there is no difficulty in deciding what to look for in assessing the results of a distance-teaching project. Say you conduct a campaign to encourage mothers to bring their babies for vaccination. The important thing to measure, obviously, is how many babies get vaccinated as a result of the campaign. However, it is generally less straightforward than this.

Suppose you publish a booklet on first aid. You hope that, because of the book, more people will do the right thing when accidents happen. But how do you measure this? You could find some people who have bought the booklet and ask them some questions about first aid, but people who give the correct answers would not necessarily do the right thing in an emergency; on the other hand, a person who cannot remember exactly what it says in the book might still have the good sense to refer to the book when an accident happens. So, test questions will not really measure the effects of the book. Obviously, you cannot create accidents in order to see how people will react to them. Could you use the official statistics of admissions to hospital? Probably not. Most accidents are small ones, such as minor cuts or burns, which never reach hospital anyway. Besides, you would not expect a few thousand copies of a book to have such a dramatic effect on the whole nation that it would be clearly reflected in hospital statistics.

In many cases, you cannot measure directly the effect that your programme has had. You have to use indirect measures instead. The guiding principle is to get as close as you can to measuring the important effect of the programme. Some examples might make it clear what I mean by 'getting as close as you can'.

Suppose you have run a distance-teaching course in bookkeeping for the treasurers of credit unions. The purpose of the course was to teach them enough about bookkeeping to manage the finances of a credit union competently. You want to find out how successful the course has been. The first idea you might have is to send them a questionnaire with questions like, 'Have you found this course useful?' Would the results of that tell you whether the course had achieved its aims? Clearly not. Almost certainly, most of the treasurers would say Yes - this might be their genuine opinion or it might be politeness - but it would still be possible that none of them had learnt enough about bookkeeping to manage a credit union's finances.

So you decide that your evaluation should contain something about bookkeeping. How about asking them to write an essay describing the five main qualities of a good treasurer? Not much better. Writing essays is very different from keeping books. Someone might write a good essay but still be a hopeless bookkeeper. (If you were about to board an aeroplane, would you be happy to know that the pilot had written an excellent essay on flying but had never actually flown a plane?)

Obviously, you have to test their competence at bookkeeping. You might write a set of test questions such as, 'What would be the compound interest after two years, at 5% per year, on a deposit of 80 dollars?' (For advice on tests, see 'Writing test questions' in Chapter 13.) This would be a lot better than the essay question. But there is still a difference between answering questions on bookkeeping and keeping books. If you could bring the treasurers together, you could do even better than this by simulating real-life tasks that the treasurers would have to perform, such as receiving a deposit from a member of the credit union or paying a bill. By 'simulating real-life tasks', I mean that you would provide the treasurer with all the things he would need to use - a receipt book, a page from a cash book and so on - and you would get him to act his part in a little play, pretending to receive a deposit from a member (or whatever) and performing the appropriate tasks exactly as he would in real life. The best evaluation of all, if possible, would be to see how they did their jobs in real life in the months following the course.

As another example, suppose you have run a nutrition campaign, encouraging housewives to cook green vegetables in a way which conserves the vitamins. You really wanted to affect their everyday cooking habits, but it is very difficult to measure this directly; you cannot walk into people's homes, pretending to be invisible, and watch them cooking. You could interview some housewives about the campaign and ask them about how they should cook green vegetables but, as with the first aid example, they might know the right answers and yet still be cooking the foods in the wrong way ('wrong' according to the campaign, that is). A slightly better approach would be to give each housewife a handful of spinach, or something like that, and ask her to cook it. If she cooked it for you in the recommended way, this would be encouraging. Of course it would still not be a guarantee that she always cooked it in that way in the privacy of her own home. But you would have discovered, at least, that she could cook it in the recommended way if she wanted to. You would still not have found out exactly what you were interested in, but you would have got closer than just by asking questions.

Asking questions or giving people a test is the most obvious way to find out what effect a course has had on them, but it is not the only way, and not always the best way. Observation (see Chapter 3) can sometimes be more reliable. Suppose you had run a course for schoolteachers to improve their teaching methods. You ask them questions as part of the evaluation but you are afraid that, though they give you the 'right' answers (i.e. the answers you want), they are not actually putting into practice the things they have been taught. The best way to find out what they are actually doing in the classroom would be to go and see.

There is a danger, of course, that the teachers would put on a special performance just for your visit. This is known in social science as the problem of reactivity; people react to being observed or tested. You are trying to observe, assess or measure their behaviour, but the very fact that you are doing this affects the behaviour you are trying to measure.

There are ways round this problem. One is to observe things that don't react. You could get some idea of the ordinary state of the classes from the pictures on the walls, the seating arrangements, the children's exercise books, art and craft materials, the condition of the school radio and so on. Another is to use records that are collected routinely for other purposes. You might use records of school expenditure on learning aids, or notes of educational outings. Another is to stay around long enough for people to get used to you, so that they stop putting on a special performance. Yet another is to talk to other people who don't have the same desire to mislead you, such as the headmaster or the children. And a fifth way is to disguise the true purpose of your visit; for instance, you might visit a school pretending to be a textbook salesman.

All of these are ways of getting at the truth about someone when the person in question would prefer that you didn't. Clearly they raise the ethical question of whether it is right to ferret out information about someone without their cooperation and perhaps against their wishes. Some researchers argue that anything is permissible in the interests of science; others argue that you should never collect any information without a person's full knowledge and consent. I myself take a middle position.

It is best to be completely open with people, if you can. In general, that is how a researcher should behave. But there are times when that is impossible.

Suppose you wanted to find out whether employers were discriminating against ethnic minorities, or whether shops were giving short measure or whether doctors were giving bad treatment. You obviously could not go along to them and ask. You would have to go behind their backs in some way to get at the truth. So I think that it is sometimes permissible for a researcher to gain information about a person without the person's knowledge and consent, but only if it is the only way to do the research and if the research is clearly in the public interest.

Someone who was trying to evaluate the teachers' upgrading course would have to decide what techniques were permissible. Myself, I'd think it was acceptable to pay a surprise visit to the school, to have a good look round and to talk to the headmaster. I'd be less happy about looking at school records or talking to the children. And I certainly would not approve of going in disguise. But it's impossible to lay down rules for this; so much depends on the particular circumstances of each case. I only want to point out that, before using any 'behind-the-back' methods, you need to think about the ethical side.

Was the effect big enough?

Finding that a piece of distance teaching has had some effect is not the end of the evaluation. Say you have evaluated a radio learning-group campaign which was intended to teach people about the national development plan. Before the campaign, people could answer, on average, three out of ten questions about the plan; after the campaign, they can answer four out of ten. The difference, let's say, is statistically significant and you are confident that the campaign can take the credit - it is most unlikely that the people would have learnt about the plan except from the campaign. Do you conclude that the campaign was a success?

Researchers are often puzzled by this sort of question. There is nothing in their textbooks of research method or statistical analysis which tells them how to answer it. That is because it is not a question of fact; it is a matter of judgement.

In describing this example, I made a point of saying that the improvement in the average score (from three out of ten to four out of ten) was statistically significant, and yet you can still ask 'Was this enough?' Statistical significance is not the sole criterion. In stating that the improvement was statistically significant, you are just saying that the difference between the 'before' and 'after' scores is genuine, and not just the result of the random errors that necessarily accompany sampling; the people really were better informed about the plan after the campaign, at least as measured by this test. But the difference is not necessarily 'significant' in the ordinary sense of the word, i.e. 'large' or 'important'. As I explained in Chapter 11, the statistical significance of a result depends partly on the size of the sample. If you had given the 'before' and 'after' tests to very large samples of people, even a tiny improvement in the average score would be statistically significant. So, even after you have established that the effect of your project is statistically significant, you can still ask, 'Was the effect big enough?'

This makes it clear that, though evaluation requires some fact-finding, there is more to evaluation than just finding facts. Having established how much effect your programme has had, you have to make a judgement about it. Because this is a matter of judgement, there is room for disagreement. The evaluator is entitled to make his own judgement about the results, but his opinion is not necessarily better than anyone else's. To return to the development plan campaign, the evaluator might consider that the improvement in average score from three out of ten to four out of ten is rather disappointing; people did not know much about the plan before the campaign and they still do not know much about it after the campaign. The campaign director, by contrast, might argue that it indicates widespread interest in the campaign and that the small but appreciable improvement in the general level of knowledge about the plan must surely be an asset to the nation's development; the campaign, in short, was a worthwhile undertaking. There is nothing in the figures themselves to say that one is right and the other is wrong.

There are many criteria that people might use in deciding whether a project was a success or not. In fact, this is a large topic that forms the subject of the next chapter. But there is one simple criterion which is the appropriate one for the kind of evaluation I have dealt with in this chapter - did the project achieve its objectives?

Earlier I stressed the value of stating the objectives in advance in a precise and testable form. You are trying to teach people about the development plan. How much do you expect the participants to know about the plan by the end of the programme? All of it? Most of it? Half of it? Just one or two of the main points from it? It would be no help if the objectives had been stated in a vague, unmeasurable way, such as 'The purpose of the campaign is to stimulate interest in the development plan,' or 'We hope to increase people's understanding of the national priorities.' But if the campaign organisers had said, in advance, that people who had participated in the radio learning groups should be able to score eight out of ten on this test, there would be no room for argument when the evaluator reported that the participants could score only four out of ten. The campaign would not have achieved its objective.

Of course, the organisers can still argue that, in retrospect, this objective was over-ambitious and that, for other reasons, the campaign was still worthwhile. But that takes us into other criteria for judging projects. In terms of this first criterion - the project's own objectives - the verdict is clear.

When you do an experiment to test a theory, you make a forecast, on the basis of the theory, about what the result of the experiment will be, and you look at the result with one question in mind, 'Was the forecast correct or not?' When you look at a distance-teaching project as if it was an experiment, you take the statement of objectives and you look at the results with one question in mind, 'Were the objectives fulfilled or not?' This is obviously an important question to ask about a project, but it is not the only one. In the next chapter I'll suggest some further questions that an evaluator ought to ask.

16 Evaluation: making value judgements

This chapter looks at other questions that an evaluation should take account of, apart from whether a project achieves its objectives.

How good were the materials? A researcher can sometimes devise a way of measuring aspects of materials.

How well did it run? You may need to design ways of monitoring a programme's efficiency.

How much did it cost? To judge whether the effects of a project justify the cost, you need some point of comparison.

Did the project have unintended effects? You cannot predict all the effects of a project. Unintended effects, both good and bad, may be more important than the intended effects.

Was the project worth doing anyway? The objectives of a project are not necessarily above criticism.

Many-sided evaluation A complete evaluation may take in many different aspects of a project.

Should the evaluator make a judgement? It is part of the evaluator's job to draw his own conclusions but he should make efforts to minimise his bias.

Insider or outsider? An evaluation may be conducted by a member of staff or by an 'outsider'. Each approach has its strengths and weaknesses.

The following quotations (all invented by me) illustrate the kind of criticisms that might occur in evaluation reports on distance-teaching projects:

> 'The correspondence course in agricultural science contains no practical work at all. This conveys to the student the false impression that agricultural science is all book-learning with no relation to the practice of agriculture.'
>
> 'Many of the radios were distributed without batteries. Consequently, many of the radio learning groups missed the first programme of the series.'
>
> 'If we divide the total cost of the literacy campaign by the number of adults who completed the course, we find that it cost 8 000 dollars for each successful student. This is unreasonably high by any standards.'
>
> 'The campaign to get farmers to use more fertiliser was so successful

> that an excessive amount of rice was produced. Consequently, the price fell and many farmers lost money.'
>
> 'For some years now, the college has maintained a good record in getting students through the General Certificate examination. But there are now many people with this certificate who are unable to find suitable jobs, and one must question whether it is beneficial, either to the students or to the country as a whole, to educate more and more people to this level.'

None of these quotations is criticising a project for having failed to achieve its objectives, not directly anyway. Indeed, the last two are criticising the projects for having achieved their objectives too well. The criticisms are, in order, about the materials, the administration and the costs, about unintended effects of the project and about whether the aims themselves were the right ones. I'll take these topics one by one.

How good were the materials?

An evaluation of a piece of distance teaching might include an assessment of the materials. This might be about the content, e.g. 'The section of the history course which deals with the colonial period gives too much attention to the European view and too little to the African view.' Or it might be about the instructional quality, e.g. 'The lesson would have been clearer if more diagrams had been used.' Or it might be about the technical quality, e.g. 'Many of the booklets had pages collated upside-down,' 'The amateur actors in the radio programme mumbled their lines.'

The researcher can give his own opinions about the materials, but his opinions are not necessarily better than anyone else's. His expertise is in research, not in history or instructional methods or printing or radio production. It is more useful if he collects, in a systematic way, the opinions of experts, or of the students, or of teachers who make use of the materials (such as the tutors in a literacy course).

To the extent that criticisms of the materials are based on fact, he can employ research techniques to see how far the criticisms are justified. If someone said that a monthly journal for farmers devoted too much space to reporting speeches by the minister of agriculture, the researcher could measure the amount of space actually taken up by these reports. When the criticism is about emphasis or bias, as in the example of a history course being pro-European, objective measurement is more difficult. But, even there, a researcher could try to make some measurement, for example by counting the number of paragraphs that presented the European view and the African view, or listing the good things and bad things that were said about European and African leaders.

I described methods of gathering opinions and assessing materials in earlier chapters (Chapters 12 and 13). I dealt with these topics at that point because it is obviously more useful to do this sort of evaluation at an early stage, while there is time for the materials producers to make improvements to the materials in the light of the comments. It may be possible to make further improvements, even after the pre-testing and initial production, if the material has a long life. A correspondence course, for instance, may last for several years and you may make improvements to it in the light of criticism from experts or from the early students.

How well did it run?

A distance-teaching project can also be judged in terms of its efficiency. If you were evaluating correspondence teaching, you would want to know whether the students were receiving the right materials at the right time, how quickly their assignments were being marked and returned to them, and so on. If you were evaluating a radio learning-group campaign, you would want to know whether the groups were meeting at the right times, whether they had all the materials they needed, whether they could hear the radio programmes, whether the group leader had received sufficient training to conduct the meetings correctly, and so on.

In the last chapter, I gave the example of how we printed monthly newsletters for an agricultural project and discovered, in attempting to assess their impact, that they had not been distributed. Another example concerns a set of word games which we produced to help children learn to read and write. As one of our experiments with these games, we gave copies to a number of primary-school teachers and asked them to use the games in their lessons. We had discussions with the teachers a few weeks later. We found, as we had hoped, that the majority had used the games successfully, but we also made some unexpected discoveries. One discovery - a pleasant one - was that several of the teachers had made extra copies of the games at their own expense, but another discovery - not so pleasant - was that some of the teachers had locked the games in a cupboard for fear that the children would spoil them.

You cannot assume that you will automatically get to hear about such problems. You have to make some effort to find out; you have to set up some system for detecting these problems. As part of an evaluation of a radio learning-group campaign in Botswana, researchers participated in group meetings and wrote their observations afterwards, and group leaders sent in reports about their group meetings. This information gave the evaluator a good picture of what actually went on in the radio learning groups. You could use similar methods when evaluating the use of distance-teaching materials in schools. If each student is studying on his own, however, you have to rely on them sending in reports or answering questionnaires. I described ways of monitoring correspondence teaching in Chapter 14.

How much did it cost?

The costs of a distance-teaching project obviously enter into one's judgement of it. If I was told that a literacy campaign had taught 1 000 people to read, and that the campaign had cost only 30 000 dollars (30 dollars per successful student), I would be impressed. But I would be less impressed if I was told that it had cost a million dollars (1 000 dollars per successful student). Distance-teaching institutions have to get money from somewhere (from governments or donor agencies or from the students) and it is fair to ask whether they provide value for that money.

Although institutions generally keep careful records of their expenditure, it is not as easy as you might think to find out how much a piece of distance teaching has cost. In Appendix 3 I describe a way of calculating the total cost of a project and I also suggest ways of converting this figure into a cost-per-something (e.g. so much per student) to make the figure easier to interpret. But let's assume that you have calculated the amount that a project has cost and have devised a suitable way of expressing it as so-much-per-something. How do you decide if this cost was reasonable or not?

If the cost is either extremely low or extremely high, the answer is obvious. Suppose you broadcast radio spots to encourage mothers to bring their babies for vaccination. You spend 150 dollars on recording and pre-testing the spots, and the broadcasting is free. A small sample survey suggests that about 100 000 mothers of young children hear these radio spots and that about 5% of these (5 000 mothers) bring their children for vaccination as a result. So a project costing 150 dollars has had some beneficial effect on 5 000 people. That makes 3 cents per person. Very good value.

Suppose, by contrast, that you are commissioned to produce a manual on weaving for a new weavers' cooperative. It turns out to be a big job, taking six months of the writer's time and a lot of work from the editor, the artist and the typist. The result is 100 copies of an 80-page manual printed in full colour. But the cooperative has few members and in fact only ten weavers ever read the manual. The total cost is 4 000 dollars - 400 dollars per reader. Absurdly expensive. (Incidentally, this example, though imaginary, is not as far-fetched as it may seem. Organisations, especially over-funded ones, often commission high quality instructional materials because they like the idea of having good-looking materials, not because anyone needs them.)

However, the answer to the question 'Is this a reasonable price to pay?' is often not so obvious. If an agricultural leaflet, costing 300 dollars, gets read by 1 000 farmers (i.e. 30 cents per farmer), is that reasonable or not? Or if a college calculates that it spends 350 dollars for every one of its students that passes an exam, is that acceptable? To help you to make a judgement about such a figure, you need something else to put alongside it; you need a point of comparison. There are several comparisons you might make.

One is to see what else you could have got for that money if you had spent it on something completely different. Say you are providing radio sets, batteries and other support materials for farmers' discussion groups. How does the cost of all this compare with buying an ox and plough for each group? Or how much seed and fertiliser could you get for the same amount? Of course, this is very rough-and-ready. You are comparing the calculated costs of a real project with the estimated costs of some hypothetical alternative. Nonetheless, it can help you to see whether the costs of a piece of distance teaching are roughly in proportion to the benefits that the students are getting from it. Try to imagine what the students would say if they were offered the choice between the distance teaching and some other project costing the same amount. If the answer, without doubt, is that they would prefer the alternative, then the distance-teaching project is probably too expensive.

Another way to assess the cost is to compare it with the cost of some other distance-teaching project that resembles it. The best comparison is with another project conducted by your own institution. Then you can make sure that the costs of the two projects are calculated on the same basis. It is more difficult to compare the costs of projects carried out by different distance-teaching organisations (especially in different countries), because organisations differ in their financial arrangements and in the way they calculate their costs. For example one organisation might have to pay to have its radio programmes broadcast, whereas another might receive this service free from the radio station. But it might be possible to make allowances for such differences, if you can find out enough about how the other organisation calculates its costs.

A third way to use costs as part of an evaluation is to compare the costs of two different ways of doing the same thing. A familiar example is the attempt to compare the costs of preparing students for some exam by distance teaching as against preparing them in the ordinary school system. LDTC took account of costs when evaluating two different ways of teaching housewives how to crochet. One way was an 'open sales' method; copies of a booklet on how to crochet were offered for sale at mother-and-baby clinics. The other was a 'group learning' method; LDTC contacted branches of the Women's Institute and offered to help group leaders to teach their group members how to crochet. LDTC used both approaches simultaneously in different parts of the country. After a few months, some of the learners were contacted (i.e. booklet buyers in the first case, group members in the second), and also a sample of women who had not received either the booklet or the teaching; they were interviewed and asked to do some crochet. In addition, LDTC kept records of the financial costs that it incurred in using these approaches.

The results showed that both methods had some success. As compared with the women who had not received the booklet or the teaching, about 40% of the booklet buyers had learnt something and almost 100% of the group members had learnt something. At the same time, the group approach was more expensive to LDTC, since several staff members spent quite a lot of time training and supporting group leaders. Calculations could be made on the basis of these results which showed that the best strategy for future programmes of that kind would be a combination of the two methods.

Did the project have unintended effects?

In distance teaching, you generally have little control over how the materials actually get used. As a result, the materials sometimes get used in ways you did not intend. An evaluator should pay some attention to this.

An example of this comes from the evaluation of the booklet on how to crochet. The book was intended for beginners, and in fact the words 'A booklet for beginners' were printed (in Sesotho) on the cover. Yet we found that a high proportion of the buyers were women who already knew how to crochet. Furthermore, a lot of these women took the trouble to teach themselves the crochet technique described in the book, in preference to their previous methods. While this was not a bad result, it was not what we had intended.

It might also be important for an evaluator to look at the side-effects of a project, rather than concentrating on the project's objectives. Sometimes the side-effects are more dramatic. For example, literacy schemes are generally undertaken with the idea that reading and writing are useful skills and that people will benefit from acquiring them. However, students of these schemes often say that the most important effect on them has been to increase their self-esteem. They feel more confident and more capable from having mastered this skill. The practical value of their new skill, enabling them perhaps to read newspapers and write letters, is a relatively minor benefit compared to this enhanced self-respect. An evaluation which ignored this aspect would be missing the main point. Similarly, a researcher who was evaluating a course in, say, English literature, should try to find out if it had increased the students' enjoyment of English literature or whether (as so often with literature courses) it had had exactly the opposite effect.

There might also be effects to look for beyond the immediate impact of the project. Distance teaching is often employed in developing countries as part of the effort to introduce modern technology into traditional societies, and the long-term effects of this process are unpredictable and not always beneficial. Bottle-feeding, for example, is a piece of modern technology that has had a decidedly harmful effect on children in many parts of the third world. Obviously, the ultimate effects of a project affect one's judgement of it. For example, an irrigation scheme might increase the yield from a stretch of land. But suppose that it also had the unintended effect of making the owners of the pumps rich and powerful and, eventually, of depriving the poorer farmers of their land. To make a final judgement on the project, you would want to draw up a balance sheet of good effects and bad effects, and you would certainly include unintended effects as well as the effects the project was aiming for.

To draw up this balance sheet of all the effects of a project, you need a research approach which is, in some ways, the opposite of the one I described in the last chapter. When looking at a project as an experiment, you try to define as precisely as possible the effect that you are interested in and you try to devise a way of measuring that particular effect. You select one aspect of the project and you focus on it. But in trying to assess all the effects of a project, including those you never expected, you don't want to confine your attention to one aspect of the project. On the contrary, you want to widen the scope of the research to make sure that you catch sight of anything that might be relevant.

There are several ways in which you might widen the scope. One is to look for other effects that the project has on the students apart from the one intended. You are more likely to find these by employing the more open-ended research techniques of observation and discussion. If that is impossible (perhaps because the students are too far away or widely dispersed), you might ask them to write, in their own words, about the effects of the project. Another way to widen the scope is to make a list of the various types of people who might be affected by the project, in addition to the students, and then to sample the opinions of each group. For example, if you were evaluating the effect of some new materials for teaching maths in schools, you might consult some teachers, principals, education administrators and parents, as well as talking to some of the schoolchildren. A third way to widen the scope is to take a longer time period. It is possible that the main effects of a project might not become apparent for a long time. If that was so, a complete evaluation would have to include some research conducted long after the project.

Was the project worth doing anyway?

When the Americans first put a man on the moon, there were, amid the congratulations, a few dissenting voices. Some people suggested that all that money could have been better spent on something like medical research or aid to the poorer countries. They were not suggesting that the project had failed to achieve its objective - it had obviously succeeded - but they were questioning the value of the entire undertaking.

When a project is evaluated solely in terms of how well it achieves its aims, the aims themselves are taken for granted. But the aims are not necessarily above criticism. Someone might argue, on ideological grounds, that a project should not have been undertaken. For example, suppose you ran a series of radio programmes to illustrate the nation's achievements since Independence; a critic might say that they were no more than propaganda for the present government. Or someone might say that a project, though not without value, was not sufficiently worthwhile to justify the effort put into it. A distance-teaching institution in a developing country might be criticised on these grounds for running a course in, say, interior decoration, instead of in something more important, such as agriculture or nutrition.

Whether or not a project's aims are worthy ones is not a question that research can answer. It is a matter of opinion. A researcher might have his own opinion, of course, and he can express it, but his personal opinion is no more important than anyone else's. An evaluation report which consisted solely of the researcher's personal opinions would not be particularly valuable. If there is controversy about the value of a project, the researcher's job is to gather the opinions of the different types of people involved and to summarise the debate in his report. If the debate centres on questions of fact, he might be able to use research to settle these factual questions and thus clarify the argument.

Many-sided evaluation

The type of evaluation I described in the last chapter is appealingly simple, in principle if not in practice. You have a clear statement of a project's objectives and you measure the project's achievements with one question in mind - 'Did the project achieve its objectives or not?' Given that the statement of objectives is precise and that the research is done competently, the answer ought to be clear.

By contrast, the sort of evaluation I have been describing in this chapter is not at all simple. You ask questions about various aspects of the project and you end up with a disorderly pile of facts, estimates and opinions. Whereas the first type of evaluation aspires to the simplicity and clarity of a scientific experiment, the second type has the complexity and fuzziness of real life.

An evaluation of the second type is like a biography, only the subject is an educational project instead of a person. And you assess an evaluation of this type in the same way that you would judge a biography. Does it cover all the important aspects? Are the factual statements accurate? Where opinions differ, are all sides of the argument presented? In short, is it reasonably complete, accurate and fair? Just as one biography of someone can differ greatly from another biography of the same person, so can evaluations differ. If it were possible for two people to evaluate the same project simultaneously, but working independently, they would probably produce quite different reports. They would begin with different preconceptions; they would focus on different aspects of the project and they would give different weight to the various opinions. (It is useful to bear this in mind when someone says, in reverential tones, 'The evaluation report said ....'; if a different person had done the evaluation, it might have said something quite different.)

The focus of an evaluation will depend to some extent on why the evaluation is being done, and for whom. If an evaluator's main task is to provide feedback for his colleagues, he will concentrate on the quality of the materials, the efficiency of the system and the effectiveness of the work, in order that the organisation can adjust or redirect its work as it goes along. If, however, he is reporting back to the people providing the funds - the government, or perhaps a donor agency - he will pay more attention to costs and to the overall value of the project. (The first kind is sometimes called 'formative evaluation' and the second kind 'summative evaluation'.)

It is not possible to give step-by-step instructions on how to do a many-sided evaluation, because the questions that one ought to ask will depend on the project and on the purpose of the evaluation. But there are two general pieces of advice I can give. The first is that much of the evaluation is based on making a comparison. One reason why people can make different judgements about the same project is that they might be comparing it in their minds with different things. People often do not make these comparisons explicit, and the resulting arguments tend to be sterile. 'The course materials were good,' and 'The course materials were poor,' appear to be contradictory opinions, but they might not be contradictory when expressed in full - 'The course materials were good, compared with other materials available in this country,' and 'The course materials were poor, compared with materials I have seen in other countries.' You can avoid this confusion by getting people to spell out comparisons explicitly.

The second point is that you should keep the evaluation in scale with the project being evaluated. Trying to answer all the evaluation questions that I've suggested would require a major piece of research and would only be worthwhile if the project itself was a big one. Evaluation is not a valuable activity in its own right. It is valuable only to the extent that it produces useful results. If you need facts in order to make a good decision, and if a bad decision might be costly, then it is worth spending money on research to provide yourself with those facts. But it is a waste of money to do research to aid decisions which are themselves of little importance.

Researchers need to be reminded of this from time to time, especially those with an academic background in social science. One piece of research raises questions for further research and it is easy to become engrossed in the subject. The researcher finds himself, like a 'pure' scientist, in pursuit of truth for its own sake. But that is not what he is being paid for. The evaluator's work itself is not above evaluation. People can ask, 'Is it efficient, effective and worthwhile?' If other people don't ask questions, the evaluator should ask them himself.

Should the evaluator make a judgement?

In the last chapter, we saw the evaluator in the role of experimenter. In this chapter, we have seen him as observer, critic and a gatherer of facts and opinions. Should he also be a judge?

Some people argue that the evaluator's job is to assemble the data (the facts, figures, opinions and so on) relevant to making a judgement on the project and then to present these in a report in such a way that readers of the report can make their own decision. The researcher (in his role as evaluator) collects the evidence; others make the judgement. The argument for this is that the researcher has expertise in gathering and analysing data, but no special qualifications for making judgements about educational projects. Besides, if the researcher expresses his own opinions in the report, this will make it difficult for others to draw their own conclusions.

This picture of the researcher's role is appealing, but it is not realistic. A researcher is not a machine that finds out everything there is to be known about a project. He has to decide which aspects to concentrate on, how much data to gather, whom to approach for opinions and what questions to ask them. Then, when he writes his report, he does not simply present a heap of facts, figures and quotations. Such a report would be unreadable. He has to select which points to make and which results to present. He has to arrange these points in a logical order and to find words and sentences in which to express them. His opinions about the project inevitably affect the way in which he does all this. A researcher who claims that his report simply presents the facts, completely without bias, is almost certainly deluding himself.

However, I do not want to suggest that the researcher's bias is a good thing. One may admit that a certain amount of bias is inevitable, but it does not follow that any amount of bias is acceptable. While there is no such thing as a report that is completely unbiased, it is obvious that some reports are more biased than others. One does not want a report to be so biased that it is completely misleading. How can the researcher cope with his own bias?

The first thing is to admit it, to recognise it and to express it. If you can say, 'I am looking at the results from this point of view,' you can then ask yourself, 'What would I think if I looked at the results from the opposite point of view?' It is difficult, but also revealing, to try to express the opinion of an imaginary opponent.

The next thing is to seek the opinions of other people, particularly those of people who are likely to disagree with you. If you are evaluating a project that you are basically in favour of, seek the opinions of someone who is likely to be hostile, and vice versa. Having listened to an opposing view, you might be able to express yourself in a less biased way in your report, or perhaps you can present both opinions, side by side.

Since, inevitably, the researcher will form his own opinions, and since these opinions will have some effect on the report he writes, I believe he has an obligation to say what he thinks in the report. Then readers can make allowances for that and they can also decide if they agree with him or not. In my experience, the chief danger for evaluation reports is not that they will have too much influence on people but that they will be completely ignored.

A question which overlaps with this one is about the researcher's freedom of speech. The researcher probably has a better understanding of the research results than anyone else and he will, naturally, draw his own conclusions from them. In fact I think it is part of his job to draw his own conclusions. However, his colleagues and, more importantly, his director, might draw different conclusions. If the report is to be typed and printed for circulation, the question arises of whose conclusions are to appear in the report. The director will not want to see a report issuing from his organisation if he disagrees with it. Equally, the researcher will not want to see a final report which incorporates his description of the research and the results and then proceeds to draw conclusions quite different from his own.

One way out of this is to adopt the policy that all reports of the organisation are official documents to be approved by the director before printing (or by the executive committee or whoever runs the organisation); in other words, the researcher is told that he does not have freedom of speech. If this policy is adopted, the researcher should be warned about it before he takes the job. The opposite policy is to allow the researcher to say what he likes, adding a sentence at the beginning of each report such as, 'The opinions expressed in this report are those of the researcher, not necessarily those of the director and staff, and they do not commit the organisation to any particular policy.' A compromise is to print different sets of conclusions in the same report. One does not often see this done, presumably because organisations do not like to admit in public that they have internal disputes, but it seems to me that, in some instances, it would be the fairest thing to do.

Insider or outsider?

Evaluation can be conducted by a researcher who is a full-time member of the organisation's staff. So far as the organisation is concerned, he is an 'insider'. Or it can be conducted by someone who has no links with the organisation - an 'outsider'. For example, a newspaper might ask an expert to review the courses offered by various correspondence colleges, rather as it commissions people to review books. Or a donor agency might arrange for an expert to visit a distance-teaching organisation for a few days or weeks and to write a report on it.

In both these examples, the evaluation is conducted for the benefit of other people, rather than for the distance-teaching organisation itself. The newspaper reviews the courses for the benefit of its readers; the expert's report is to help the donor agency make its decisions. But the distance-teaching organisation, if it gets a copy of the review or report, can also benefit from it. If it thinks the criticisms are fair, it can take action to improve its work. From this comes the idea that if an organisation felt that an outside opinion would be useful to it, it might itself commission an outsider to do an evaluation.

People debate at length whether it is better to have evaluation done by an insider or an outsider. In fact, the two are, simply, different, and each has its value. Indeed, they are complementary, by which I mean that where one is strong, the other is weak, and vice versa.

The insider's great strength is that he understands the project in detail. He knows the people involved; he knows how the organisation works; he knows about the country; he knows the problems that his colleagues have to face. He can take account of all these things when assessing the successes and the failures of the project. Being a member of the organisation, he can sometimes get his colleagues to help in the evaluation. For example, he might get them to do their fieldwork in a particular order (as in the example of the film vans, on page 211), or to collect certain items of information as part of their routine.

Unfortunately, this close involvement in the project is also the insider's weakness. If you work in an organisation for more than a few weeks, you begin to take many things for granted. You assume, without realising it, that certain things cannot be changed. So an inside evaluator might pay no attention whatever to certain important aspects of a project because he has ceased to notice them.

A further problem with close involvement is that an insider cannot be impartial. If you work on a project, even one that you have reservations about, you become committed to it and defensive about it. So an inside evaluator can generally be assumed to be in favour of the work that his colleagues are doing. Like them, he wants the project to succeed; he wants his evaluation results to show that it is succeeding. Perhaps the future funding of the organisation (and therefore his own salary) depends on the success of the project.

At its worst, this can mean that he purposely distorts the evaluation results. For example, he can write questions in a questionnaire in such a way that respondents will tend to give favourable answers. He can arrange the tables of results to make the project look good. He can simply ignore, or destroy, research results which are not to his liking. Honest researchers will not do such things, of course, but researchers, as a class, are not necessarily more honest than any other set of people.

However, even honest researchers find it hard to prevent bias creeping into their evaluations. Say that a small piece of evaluative research suggests that a project was much less successful than had been hoped. If these are not the results you wanted, it is natural to seek alternative explanations for them, or you might repeat the research with a different sample in the hope of getting different results. By contrast, if the small piece of research suggests that the programme was successful, you will be inclined to accept the results at face value. In the last section I described ways of trying to reduce bias, but the best you can do is to minimise bias; you cannot eliminate it completely.

The outsider's strengths and weaknesses are the opposite of these. He does not have the insider's detailed understanding of the project. If he is a visitor from another country (as is often the case) he might have very little knowledge of the country and the culture. So he runs the risk of making criticisms or suggestions which, though reasonable in the context of his own culture and experience, are impractical or irrelevant to the project in question.* And he cannot make use of any research designs or data collection that require close collaboration with the project staff. (If he stays long enough to achieve this collaboration, he might just as well be an insider, even though he may be paid by a different organisation, since he will start acquiring the insider's weaknesses.)

On the positive side, is he free from bias? Clearly, he does not have the natural commitment to the project that the insider has, so, to that extent, his judgement is less likely to be distorted. It is easier for him to say harsh things about a project if he thinks they are justified. But I argued earlier that no one is completely free from bias, and this includes outside evaluators.

An outsider does not arrive with a completely open mind. If his visit has been commissioned by some agency - perhaps a donor agency - they will have briefed him in advance. They have commissioned an evaluation for some particular reason and they will convey this to him. Perhaps they have heard rumours that their money is being misspent and they want him to look closely at the project's accounts. Or perhaps they think the project is being manipulated for political ends and they want him to be on the lookout for signs of this. In addition to this, the outside evaluator has his own preferences and prejudices, and he brings these with him. Suppose the project is using educational television. If the outsider has, himself, run a famous educational television project and writes articles promoting this teaching method, he is likely to think well of it. If, on the other hand, he is a person who is known to be suspicious of educational technology, he is likely to be critical of the project. Agencies are aware of this, of course, and might appoint an evaluator who, they think, will tell them what they want to hear.

The outsider's main strength is not that he is free from bias. It is that he notices things that the insiders take for granted. To give a minor example, I once visited a distance-teaching unit which, because of the shortage of office space, was divided into two sections about 3 km apart. LDTC had always been lucky enough to get offices which could accommodate all the staff in one place, so I asked the staff of this unit whether the splitting of the organisation created problems. They all assured me that it created no problems at all. Later on, I asked if I could borrow one of their course books. The director said I could, but that the books were kept at the other place. I went to the other place and the storeman told me he could not give me a book without a form filled in by the director. We phoned the director who, by this time, had gone home. I did not get the book till the following day. The staff of this organisation had got used to this arrangement and, when they said it created no problems, they meant that they were still able to function adequately. It was clear to me, as an outsider, that it made them less flexible and less efficient than they would be if they were all in one place.

People working in an organisation tend to get engrossed in the day-to-day running of the work. They might be dimly aware that the whole project is moving off course or that the organisation's future is not as clear as it should be, but they put these worries to one side while they concentrate on more pressing problems. An outsider, without the day-to-day problems, can more easily take a long-term view. What plans have been made for the continuation of the project when the present funds run out? Who will take over the senior posts when the present people leave, and are they being trained for it?

---

* You can minimise this weakness by inviting an outsider who is at least familiar with the type of work you are doing. For example you might borrow a researcher from another distance-teaching organisation for a fortnight or a month, to make an assessment of your organisation's work.

I said earlier that the strength of the inside evaluator is that he understands the project in detail. For an outsider, it can be an asset if he misunderstands the project. At LDTC we had begun producing booklets on practical topics as a way of offering useful instruction to rural people. In our eyes, the booklets were an educational medium. We had a visit from a man whose background was in book publishing. He looked on our booklets scheme as a publishing venture. We had to admit that our booklets scheme (of which we were rather proud) looked amateurish when considered from the viewpoint of a publisher. How many titles per year could our market absorb? We hadn't thought of it like that. What about the author's copyright? It hadn't crossed our minds. Why did we produce 24-page booklets instead of the more economical 16-page or 32-page ones? We didn't know about that. And so on. It was not that he was more intelligent or generally more knowledgeable than us. It was just that he looked at our scheme in a different way, and this was very useful.

To sum up, much of the research that I have been describing in this book requires a researcher who is on the staff. Only then can research be integrated into the regular activities of the organisation. The inside researcher brings to evaluation an intimate understanding of the organisation and the familiarity that comes from daily contact with a project. An evaluation by an outsider is not a substitute for this, nor is it necessarily more objective, though it might be. The outsider's main advantage is that he brings a fresh pair of eyes. He can ask important questions that the staff had not thought of. He can make people think about things they had taken for granted.

17 After completing the research

Feeding results back to the people who are supposed to take notice of them is a process that requires careful handling.

Collaborating with non-research colleagues *Integrating research into the work of the organisation is part of the researcher's job.*

Writing research reports *Discuss the interpretation of results with colleagues before writing the full report. Be diplomatic in writing it and make it as readable as possible.*

Conflict between action and research *There is no general rule for giving priority to research or action when they conflict.*

The value of research *There is no guarantee that people in distance teaching will take notice of research results but their work is more likely to be effective if they do.*

The kind of research I have been talking about in this book is intended to produce results that are useful. The basic idea is that the researcher discovers facts, and then his colleagues in the distance-teaching organisation modify their work in the light of these facts in order to make it more effective. That is, as I say, 'the basic idea'. In practice, it is not so simple.

In Chapter 2 I described a number of ways in which people can commission research, supposedly out of a desire to be guided by the results, but in fact without any intention of making changes in their work. If a piece of research is not directed towards some policy question, even at the outset, then the results are not likely to affect what the organisation does.

I have also mentioned the danger of designing an over-ambitious research project whose results arrive too late. The decisions that the research was meant to guide have already been taken by the time the results arrive. Or, even if no formal steps have been taken, the main lines of a plan have been formed; people's minds are no longer open to new suggestions.

Even when the research results are useful and timely, they do not necessarily influence action in a straightforward way. For one thing, there is room for disagreement about the policies to be adopted in the light of research results. Suppose that an organisation is broadcasting a weekly programme giving fifteen minutes of agricultural advice. The research department conducts a survey of the radio audience and discovers that hardly anyone is listening to this programme. What is to be done? One staff member might conclude that the programme is a waste of resources, so it ought to be scrapped. Another might conclude that one weekly programme of fifteen minutes is not enough; what is needed is three weekly programmes of thirty minutes each, presenting agricultural advice in the form of a drama with lots of music. So research results do not lead in a simple way to the adoption of a particular policy.

Secondly, people are human. This obvious fact is often ignored by scientists, even social scientists. People are sensitive to research results that imply criticisms of their work. An experienced geography teacher who has spent two months producing a first volume of a course on secondary school geography does not want to be told that students find it confused and boring. A project director who has spent two years working hard on a big campaign does not want to be told that it has had very little effect; nor does the donor agency who sponsored it, or the senior civil servant who approved it. As well as their personal feelings, people's status, job security and promotion prospects might be bound up in a piece of work. You cannot expect them to make a cool and objective appraisal of research results that touch on a matter so close to their hearts.

In addition to this, many people simply do not understand research results, especially when the results are expressed in figures. They will be mildly perturbed by a column of percentages, decidedly intimidated by semi-technical expressions like 'random sample' and 'control group', and totally mystified by a standard error or a significance level. What they cannot understand they will simply ignore.

The consequence of all this is that research results are not necessarily used to modify policy in an enlightened way. If a research report contains unpalatable findings, it might be ignored or even censored or suppressed. People will exaggerate the parts they like, ignore the parts they dislike, or just misquote results in order to support their case. The researcher cannot prevent this happening, but, if he recognises the problem from the beginning, he can take steps to minimise it.

Collaborating with non-research colleagues

The first thing is that the researcher should look upon himself as a member of a team. Close co-operation between the research department and the rest of the organisation does not just happen on its own. The researcher must make it part of his job to show his colleagues how they can make use of research and to involve them in the design of research projects. As I emphasised in Chapter 2, thorough discussion of a piece of research is essential at the earliest stage. Both the researcher and the colleague who is commissioning the research must try to see, as clearly as possible, what the purpose of the research is, what kind of results it might produce, and what the non-research colleague might do with the results. Far too often it is only when he has the final report in front of him that the non-research colleague asks, 'What am I to do with these results? How does this help me?' By then it is probably too late. The researcher, in the early discussions, should give him some idea of what to expect at the end.

Having established what it is that his colleague wants to get out of the research, the researcher should devise a research design that will produce the necessary information. He should assess the level of accuracy that will be adequate; there is no point producing results accurate to 5% if results that are accurate to 15% will do. He should bear in mind the time it will take; obviously, he must produce results in time for his colleague to use them. In general, he should also favour a simple research design over an elaborate one, since the important thing in the end is that his colleague, not just the researcher himself, should be able to understand the results.

How far the non-researchers can be involved in carrying out the research depends on the project. Their involvement can be valuable, especially in pretesting. For example, if a researcher tests a draft correspondence lesson with a few students and reports the results to the writer, the writer might take account of them or he might not. But if the writer himself attends the pretest, observes the students struggling with his draft lesson and discusses their questions with them, he is more likely to modify his draft. Even if the non-researcher is not involved in conducting the research, he should understand fully what is being done so that he can make his own assessment of the reliability of the results. Again, it is up to the researcher to explain to his colleagues what is being done.

There is a problem in this close collaboration, namely the problem of bias. If the researcher is co-operating closely with a colleague, it is more difficult for him to be critical or to ask uncomfortable questions. He will tend to interpret results in the most favourable way. This is particularly true in evaluation. There are steps one can take to reduce the amount of bias - I described some in the last chapter - but it is naive to suppose that one can eliminate bias completely. People using the research results have to try to make allowances for this.

If collaboration runs the risk of producing research that is somewhat biased, lack of collaboration runs the more serious risk of producing research that is irrelevant or ignored. If the researcher fails to make efforts to adapt his work to the needs of his colleagues, the research unit will tend to become isolated, and viewed by the rest of the staff with anxiety and distrust. People will not invite the research department to help them solve their problems; in fact they will tend to conceal their work from the researcher. People will not believe the research results, or they will find excuses for taking no notice of them; perhaps they will not read the research reports at all.

If a researcher wants his work to be taken account of and acted upon, he has to be a diplomat as well as a social scientist. He has to establish and maintain good relations with other staff members, even though the research results might sometimes imply criticism of their work. As a diplomat, he must be realistic about staff politics. Whose interests are affected by a particular piece of research? Is someone trying to prove something? If so, what are they trying to prove, and for what purpose? How might the research results affect decisions? Does anyone stand to lose or to gain from these decisions?

I am not suggesting that the researcher should allow his work to be dominated by staff politics. For example, it would obviously be wrong for him to fabricate results in order to support a colleague whose opinions he happened to agree with. But staff politics will certainly affect the way in which the results are used, and the researcher who is aware of the politics can devise his tactics accordingly, by making sure that he discusses the research with the right people at the right time, for example, or by finding a way to present uncomfortable results without causing needless offence.

Writing research reports

The end product of a piece of research is usually a written report. This is not always so. If you conducted a quick pre-test of a few sketches, you might just discuss the results with the artist and leave it at that, without going to the trouble of writing out a report. But, in general, it is worthwhile to make the effort of writing a report. It forces you to clarify your thoughts; you have to decide which results are the most important and what conclusions should be drawn from them. It provides your colleagues with a document on which to base their discussions and meetings; they might draw different conclusions from the results, but at least they all have the same figures set out in the same way. The report is something to refer back to in the future; if someone says, rather vaguely, 'Last time we did this sort of thing, we found such-and-such,' you can get out the report and check on what the results actually were. The pile of research reports that accumulates over the years provides a useful introduction for new members of staff; by reading the reports, they can profit from the experience of their predecessors and avoid repeating their mistakes. Finally, the written report is something you can send to other agencies who are interested in your work, or to people doing similar work in other countries.

In Chapter 9, I dealt with the technical content of research reports: this section is about the diplomacy of report writing.

In analysing results, a researcher might arrange the data in dozens of ways. He collects a large pile of figures and crosstabulations. But he does not present all these tables in the report. If he did, the report would be unreadable. He has to decide what the main findings are and then to select those tables that show these findings. In short, he has to decide what he is going to say in his report. At this stage it is a good idea for him to present some of the results to his colleagues and to ask them what conclusions they draw. This gets them thinking about the results and it enables the researcher to incorporate their opinions in his report. In addition, they are more likely to read the report carefully if they feel they have made a contribution to it. A further opportunity for getting other opinions on the results is when the researcher has written a draft of his report; he can circulate it among his colleagues for comment and criticism.

While these points apply to the whole of the report, they are particularly relevant to the 'conclusions and recommendations' section. It might be sensible to circulate a draft of the earlier parts of the report first, and only to write the last section after collecting people's opinions on the earlier parts.

Interpreting the facts - expressing them in a certain way, drawing conclusions from them - is a process that needs handling carefully. It is the interpretation, rather than the facts themselves, that hurts or pleases. The fact itself might be, 'Of the 100 students who enrolled in maths at the beginning of the year, 50 had submitted a worksheet in the next twelve months and 50 had not.' It makes a big difference whether you say (or imply), 'A drop-out rate of 50%; this is terrible; we must reorganise the student counselling service,' or 'A 50% drop-out rate is roughly what you'd expect in this kind of work,' or 'Half the students are managing to get started on the maths course; that's unexpectedly good when you consider all the difficulties.'

It is in evaluation reports that these problems of diplomacy are most acute. Try to give as much stress to the positive aspects as to the negative ones. Evaluation reports tend to concentrate on what went wrong, and I think this is because people find it more natural to comment on faults. If your train arrived ten hours late you might write a letter of complaint to the railway company, but you probably wouldn't think of writing a letter of thanks just because it arrived on time. One feels that, when an organisation functions properly, it is only doing what it is supposed to do; one only comments if it falls short in some way. When writing evaluation reports you should resist this tendency and try to balance harsh words with kind ones.

A particular problem arises if the project being evaluated was undertaken in conjunction with another agency, especially if the other agency is not accustomed to having its activities evaluated. In such cases, you have to make it clear in advance to the other agency that the project will be evaluated. You also have to make a point of inviting them to give their interpretation of the results before you finish your report, and you have to be more than usually diplomatic in writing it. It might be tempting to say, 'We did our part satisfactorily, but they made a mess of theirs,' but it would not be tactful, even if it were true. Evaluation, after all, is intended to lead to projects being more successful in the future. If the evaluation only makes people upset and unco-operative, it has failed in its purpose.

Earlier in this chapter, I said that one of the advantages of a written report is that you can send copies of it to other agencies who are interested in your work. I must add a word of caution to this, however, which is that, when you are writing the report, you must bear in mind the people who might eventually be reading it. For example, an evaluation report which listed the shortcomings of a project fully and frankly might be very helpful to staff members in deciding policy changes, but, if it was given to outsiders, it might convey a misleading impression that the organisation was grossly incompetent. This could have serious consequences if the outsider was, say, a senior civil servant or the representative of a donor agency. So you should decide before you write the report whether it is for general circulation or only for the eyes of your colleagues. If it is to be restricted to staff members, you must make sure that it is clearly labelled as an internal document.

All of this diplomacy, I need hardly say, will be wasted unless the report itself is readable. The basic purpose of the report is to communicate the findings to the readers. The readers will generally not know much about research or statistics, so the report must be clear and simple. It should also, if possible, be short. Select the important points and make sure that they stand out. Include minor findings if you think they might be useful but don't give the reader more than he can take in. Keep technicalities out of the main body of the report; put them in a separate section, or perhaps in an appendix. Use tables and diagrams but make sure they are clear and don't include too many. Most important of all, say what you have to say as briefly and as plainly as you can.

A university training can be a handicap for the researcher when it comes to writing clear reports. In courses of social science at a university, one writes research reports primarily to show the teachers or examiners that one has mastered the technicalities of the subject, so students make a point of using technical terms and showing off their statistics. Researchers who have acquired this habit should try to get rid of it. Take a sentence like this: 'The correlation between student age and worksheet completion rate (Pearson's product-moment coefficient of correlation) was found to be -0.17, a level well short of statistical significance..' This would be meaningless to the great majority of readers and should be rephrased along these lines - 'We looked to see whether the younger students sent in their worksheets faster than the older students but we found no evidence of this.' Better still, ask yourself whether it is worth reporting at all.

When writing a report, there is always a fear in the back of your mind that if you make the report easy to read, some readers may imagine that the research was easy to do. Such readers do exist and it is tempting to make a report needlessly technical in order to impress them, but you must resist this temptation. The readers you ought to be writing for are those who simply want to understand what you did and what you found; the more plainly you express yourself, the more grateful they will be.

Having made the effort to write a report, it is worth making some extra efforts to encourage people to read it. One way to get your colleagues to read it is to give a seminar on it; you invite the key people to the seminar, warn them that the report will be discussed and tell them that they will be expected to have read it beforehand. Even if they still do not read it, you can put across the main points at the seminar. As for people in other agencies and other countries, you obviously cannot do much to encourage them to read your report. But you can at least make sure that they get a copy, by building up a mailing list of interested agencies and sending them copies of research reports (the ones for general circulation, that is) as a matter of routine.

Conflict between action and research

In the long run, there is no conflict between action and research. The basic function of a distance-teaching institution is to teach; action, that is to say, is the important thing. Research is done in order to make the action more effective, and that is what everyone wants. But there can be conflicts in the short run.

Suppose that an organisation has decided to do a six-month trial of a weekly radio programme. It becomes apparent, after only the first five weeks, that no one is listening to the programme, so the producer decides to make drastic changes to the style and timing of the programme. The researcher, however, has designed his evaluation on the assumption that the original style of the programme would be given the full six-month trial; he protests that he cannot do a careful evaluation of the programme if the producer keeps changing it.

In this case, I myself would take the side of the producer. No doubt the evaluation results would be more accurate and more reliable at the end of the full six months, but it is already obvious what the gist of the evaluation would be, namely that the programmes had failed to educate anyone because nobody had listened to them. It would be more useful to spend the remaining months of the trial conducting an evaluation, albeit a less thorough one, of the new-style programme rather than amassing data to prove that the original type of programme had been a failure.

But there are also times when research should take priority over action in a short-run conflict. Sometimes a researcher wants to conduct a baseline survey before an education campaign begins. He needs this baseline data in order to assess, later on, the impact that the campaign has had. The campaign organisers, however, might be very keen to begin their campaign. They have prepared the educational materials and they are eager to put them into circulation.

In this case, I would take the side of the researcher. A baseline survey may be a crucial part of the evaluation of a campaign; without it, the evaluation would be little more than guesswork and wishful thinking. It may be important enough to justify delaying the start of the campaign by two or three months in order to allow the baseline survey to take place.

I have made these examples fairly clear-cut. Real life, of course, is more complicated. But the point I am making is simply that there is no general rule for resolving these short-run conflicts; you cannot say that action should always take priority, or that research should. You have to weigh up the importance of the action and the research in each cas

The value of research

I hope I have shown that research can be useful; it can help a distance-teaching organisation to do better distance teaching. But I do not want to make too big a claim. Not all research is useful. Quite a lot of research that gets done is a waste of resources, because the purpose of the research is not thought out clearly enough at the outset, or because the research is designed or conducted so badly that it fails to answer the questions that it was intended to answer. Even good research is not necessarily made use of, because the people who are supposed to act on the results do not get to hear about the results, or do not understand them, or choose to ignore them. So the researcher should not assume that his research will automatically put his organisation on the right road.

But in the absence of research, people in distance teaching put money and effort into activities about which they have no information or, worse, wrong information. Some resources may be wasted on useless research, but far more are wasted on misguided effort. When the students are at a distance, it is easy to delude yourself into thinking that things are different from the way they really are. It is all too easy to print thousands of leaflets that no one can understand, to broadcast hours of radio programmes that no one listens to or to despatch packets of correspondence courses that no one looks at. It is quite possible for a distance-teaching unit to carry on like this for years, vaguely imagining that its efforts are bringing enlightenment to thousands. Research can put facts in the place of these delusions. Research cannot guarantee that people will adopt the best policies, but it can bring a bit of realism to their thinking.

Appendix 1 Statistical procedures

In Chapter 11 I described some statistical concepts. Here I explain how to perform the calculations. I begin with a section to help you decide which pages you need to read. (Don't try to read this appendix straight through from start to finish.) For most of the calculations you will need a calculator. An ordinary pocket calculator will do, so long as it has a square-root button (√).

Many readers will be content simply to work through the steps of the calculations and to accept that the results mean what they are supposed to mean. But some may wonder how these formulas were invented. To put the question in another way, how is it that the results of these calculations tell you what they are supposed to tell you? If that question crosses your mind, read Appendix 2 in which I sketch the statistical theory underlying these procedures.

Which parts of Appendix 1 do you need to read?

1. Do you know how to read statistical formulas? For example, can you understand this:

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

If you can't, then start by reading pages 242 to 247.

2. If you have a result from a random sample (e.g. '35% of the farmers we interviewed were using the new type of seed') and you want to calculate the standard error (for setting confidence limits) see page 248. You should also read this section if you are going to do a sample survey and you are wondering how many people to interview.

3. To calculate a standard deviation, see page 259.

4. If you want a test of statistical significance, are your results in terms of category variables or measurement variables (see page 151), i.e. which of these do they resemble?

Category variable (Passed/Failed)

Groups	A %	B %
Passed	62	53
Failed	38	47
Base total	(119)	(137)

Measurement variable (test score)

Groups	A	B
Mean score (out of 10)	8.6	7.3
Number in group	(119)	(137)

If your results are in category form, try the test of the difference between proportions (page 250) or the chi-squared test (page 252). The chi-squared test is better if you have a table larger than two-by-two.

If your results are in measurement form, read 'Dependent and independent scores' on page 260.

5. If you want some other statistical procedure, consult a statistics book - Appendix 4 (page 297) gives some references - or consult a statistician.

Formulas: 'how-to-do-it' instructions in shorthand

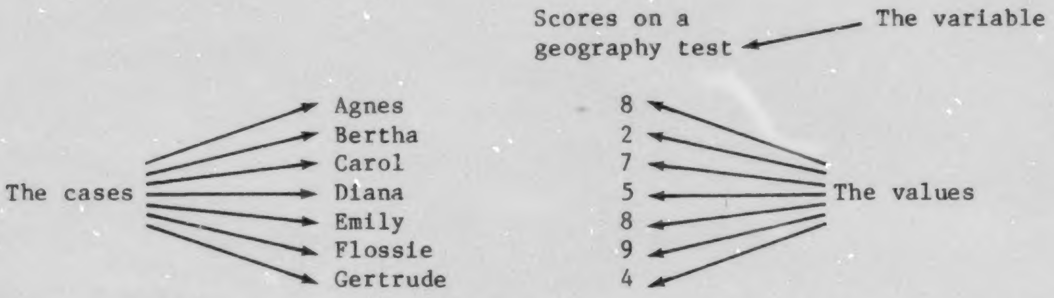
Formulas are perhaps more responsible than anything else for making people nervous about mathematics. To people who understand them, formulas* are simply a kind of shorthand - a quick way of writing down complicated things; to people who don't understand them, these complex arrangements of Greek letters and mathematical symbols can be mysterious and intimidating.

I find it easiest to understand a statistical formula as a condensed set of 'how-to-do-it' instructions. So long as you follow the instructions step by step and in the right order, you reach the correct solution. The symbols are the instructions, so you need to know what they mean, and you also have to know what order to take them in.

As an example, take the instructions for calculating the average (the mean). Let's say you've done a survey of poultry owners; you know how many birds each person owns and you want to calculate the average poultry holding. You add up all the numbers and divide by the number of holdings. This can be expressed by the following formula:

$$\bar{x} = \frac{\Sigma x}{n}$$

Before I explain this formula, do you remember what the terms 'case', 'variable' and 'value' mean? I explained them on page 151 and used this diagram:



Returning now to formulas, the letters 'x', 'y' and 'z' are often used in statistics to stand for the variables. (They are the ordinary letters x, y and z from the English alphabet.) 'Σ' is the capital letter sigma from the Greek alphabet and it is known as the 'summation sign'. 'Σx' simply means

---

* Some people use the Latin plural of 'formula' - 'formulae', pronouncing the 'ae' to rhyme with the 'ee' in 'see'. Others, like myself, use the plural 'formulas'.

'The total of all the values of the variable'. In the example I've just given, the variable represented by x is the test score, so 'Σx' means 'the total of all the scores':

$$\Sigma x = 8 + 2 + 7 + 5 + 8 + 9 + 4$$

In statistical formulas, the letter n is used to mean 'number of cases'. In the example of the test scores, the cases were schoolgirls and there were seven of them. So, in this example, $n = 7$.

The formula for the average puts these two expressions together:

$$\frac{\Sigma x}{n}$$

This means 'the total of all the values divided by the number of cases' or you can read it as a set of instructions that says, 'Add together all the values of the variable (that's the Σx part) and then divide by the number of cases.' For the schoolgirls' geography scores, this means:

$$\frac{8 + 2 + 7 + 5 + 8 + 9 + 4}{7}$$

$$= \frac{43}{7}$$

$$= 6.1 \text{ (rounded to one decimal place)}$$

The only bit of the formula that I haven't explained is the '$\bar{x}$'. This is read aloud as 'x bar' and it is the statistical symbol for the average (the mean) of the variable x.

To sum up, the formula for the average is:

$$\bar{x} = \frac{\Sigma x}{n}$$

You can read this as a statement which says:

2 the values of the variable added together...

1 The mean is... $\bar{x} = \frac{\Sigma x}{n}$

3 and divided by the number of cases.

Or you may prefer - as I do - to read it as a set of instructions which says:

2 you add together all the values of the variable,

1 To calculate the mean, $\bar{x} = \frac{\Sigma x}{n}$

3 and then you divide by the number of cases.

As a slightly more complicated example, suppose you gave some students a test in agricultural science, then you gave them a lesson in that subject, and finally gave them the same test again. For each student you have two scores - one from the test before the lesson and one from the test after - and you want to calculate the average improvement in their scores. For each student you subtract the before-score from the after-score to see how much he has improved, then you add together these amounts-of-improvement and divide by the number of students. You could express this, if you wanted, in the following formula:

$$\overline{(y - x)} = \frac{\Sigma(y - x)}{n}$$

For each case (each student), you have two variables (two test scores), and I've used the letter x to stand for the score on the test before the lesson and y for the score after the lesson. '(y - x)' is the difference between the two scores, or, in other words, the amount of improvement. The calculations might look like this:

	Before score x	After score y	Improvement (y - x)
Alfred	7	9	2
Benjamin	4	5	1
Cuthbert	3	6	3
David	4	7	3
Edward	6	8	2
Frederick	7	10	3
		Total $\Sigma(y - x)$	14

$$\frac{\Sigma(y - x)}{n} = \frac{14}{6} = 2.3 \quad \text{(to one decimal place)}$$

Reading the formula as a set of instructions, it says:

3 then you add them up,

1 To calculate the average improvement in score,

$$\overline{(y - x)} = \frac{\Sigma(y - x)}{n}$$

2 you subtract the 'before' from the 'after' for each student,

4 then you divide by the number of students

Formulas: symbols and the order of operations

Most people are familiar, from their schooldays, with the symbols for the four basic operations of addition, subtraction, multiplication and division:

+ - × ÷

Only the first two of these commonly appear in formulas. Instead of writing '$x \times y$' for 'multiply x by y', you generally just write 'xy'; and instead of '$x \div y$' for 'divide x by y', you write x/y. The following table shows what these expressions mean, for various values of x and y; e.g. if x is 3 and y is 2, $x + y$ is 5:

x	y	$x + y$	$x - y$	xy	x/y (also written as $\frac{x}{y}$)
3	2	5	1	6	1.5
4	10	14	-6	40	0.4
95	65	160	30	6175	1.46 (to two decimal places)

Squares and square roots are often used in statistical formulas. x^2, which is read as 'x squared', means $x \times x$. $\surd x$, the square root of x, is that number which, when multiplied by itself, gives x. The square root of 9, for example, is 3, because $3 \times 3 = 9$. This table gives the values of x^2 and $\surd x$ for various values of x:

x	x^2	$\surd x$ (also written as $\sqrt{x}$)
9	81	3
25	625	5
61	3721	7.81 (to two decimal places)

Three more symbols which often appear in statistical formulas were all used in the formula for the mean in the last section; they are n, $\bar{x}$, and Σ. 'n' stands for the number of cases. The bar sign in $\bar{x}$, or $\bar{y}$ or $(\overline{y - x})$, stands for the mean of the expression under the bar, i.e. the mean of the variable x, or y or $(y - x)$. And 'Σ' means 'the total of'; 'Σx' for instance means 'the total of all the values of x'. Suppose you had the scores of four schoolchildren on a maths test, and you used 'x' to stand for the test scores:

	x	x^2	$\surd x$
Ann	4	16	2
Bill	9	81	3
Cathy	16	256	4
Derek	25	625	5
Total	54	978	14

In this example Σx would be the total of the values of x i.e. 54. 978 is the total of the values of x^2, i.e. Σx^2; and 14 is the total of the values of $\surd x$, i.e. $\Sigma \surd x$.

As I've said, you can read a formula as a set of instructions telling you how to calculate something. If you understand the symbols you can see what steps you have to take, but you also need to know what order to take them in. With most formulas, the order in which you do things is important. Suppose you have three numbers, represented by x, y and z, and a simple formula:

$$x - y + z$$

This might mean 'Subtract y from x and then add z,' or it might mean 'Add y to z and subtract that from x.' The difference is important. Let's say x is 10, y is 6 and z is 3. These are the two ways you might tackle it:

$$\begin{array}{ll} 10 - 6 + 3 & 10 - 6 + 3 \\ = 4 + 3 & = 10 - 9 \\ = 7 & = 1 \end{array}$$

The way that a formula is laid out usually makes it clear what order to do things in. Take these two examples, supposing that $x = 2$ and $y = 3$:

$\frac{x}{y} + 1$	$\frac{x + 1}{y}$
Divide x by y and then add 1.	Add 1 to x and then divide by y.
e.g. $\frac{2}{3} + 1$	e.g. $\frac{2 + 1}{3}$
$= 0.67 + 1$	$= \frac{3}{3}$
$= 1.67$	$= 1.0$

If the layout does not make it clear, a general rule is that you do multiplication or division before you do addition or subtraction. For example:

$$x\,y + z$$

means 'Multiply x by y and then add z.' Suppose x is 2, y is 3 and z is 4, then:

$xy + z = 2 \times 3 + 4 = 6 + 4 = 10.$

Brackets play an important part in telling you what order to do things in. The rule is that you calculate what's inside the brackets first. Brackets could be used in the earlier example in two ways:

$(x - y) + z$	$x - (y + z)$
Subtract y from x and then add z.	Add y to z and then subtract the result from x.
e.g. $(10-6) + 3$	e.g. $10 - (6 + 3)$
$= 4 + 3$	$= 10 - 9$
$= 7$	$= 1$

As another example of the use of brackets, consider again the four schoolchildren's test scores. Adding up the column headed 'x^2' gave the total of 978, represented by Σx^2. This could also be written, with brackets, as $\Sigma(x^2)$. If the brackets were placed differently, as in $(\Sigma x)^2$, this would mean something else; it would mean the square of Σx. Σx was 54, so $(\Sigma x)^2$ means 54 × 54 = 2916:

$\Sigma(x^2)$	$(\Sigma x)^2$
Square each value of x and then add them up.	Add up the values of x and then square the total.

Here are a few more examples to show how the positioning of symbols, especially the brackets, indicates the order in which to do the calculations:

	x	$x - 1$	x^2	$\sqrt{x}$
Cases A	4	3	16	2
B	6	5	36	2.45
C	10	9	100	3.16
	$\Sigma x = 20$	$\Sigma(x - 1) = 17$	$\Sigma x^2 = 152$	$\Sigma\sqrt{x} = 7.61$

Note that $\Sigma(x - 1) = 17$, but $\Sigma x - 1 = 20 - 1 = 19$.

$\Sigma x^2 = 152$, but $(\Sigma x)^2 = 20 \times 20 = 400$.

$\Sigma\sqrt{x} = 7.61$, but $\sqrt{\Sigma x} = \sqrt{20} = 4.47$.

	x	y	$x + y$	$(x + y)^2$	x^2	y^2	$x^2 + y^2$
Cases A	4	2	6	36	16	4	20
B	6	7	13	169	36	49	85
C	10	5	15	225	100	25	125
Totals	20	14	34	430	152	78	230

Note that $(x + y)^2$ is not the same as $x^2 + y^2$.

	x	y	x^2	xy	$(xy)^2$	x^2y	x/y	$(x/y)^2$	x^2/y
Cases A	4	2	16	8	64	32	2.00	4.00	8
B	6	7	36	42	1764	252	0.86	0.74	5.14
C	10	5	100	50	2500	500	2.00	4.00	20
Totals	20	14	152	100	4328	784	4.86	8.74	33.14

$\Sigma(xy) = 100$, but $\Sigma x \, \Sigma y = 20 \times 14 = 280$

$\Sigma(\frac{x}{y}) = 4.86$, but $\Sigma x/\Sigma y = \frac{20}{14} = 1.43$

One final point that I should mention for readers who are out of practice with their mathematics is that if you multiply a minus number by another minus number, the answer is a plus number. For example $(4 - 7)^2 = -3 \times -3$, and this makes 9, not -9.

The standard error

The way you calculate the standard error depends on what type of random sample you have. First I'll give the formula for the standard error from a simple or systematic random sample (these terms are explained on page 41).

Suppose a college wants to know more about the circumstances in which the students are studying at home. It has 4300 correspondence students on its books and has drawn a systematic random sample of 172 by taking every twenty-fifth name after a random start. The college has contacted all of these and has found that 81% are studying in the evening by candlelight.

To calculate the standard error, which I will abbreviate to 'SE', the formula is as follows:

$$SE = \sqrt{\frac{pq}{n - 1}}$$

In this formula, P is the percentage in question, q is (100 − p) and n is the number of cases. In the example of the students studying by candlelight, p is 81%, q is (100 - 81) = 19%, and n is 172. So the standard error for this result is:

$$\begin{aligned} SE &= \sqrt{\frac{81 \times 19}{172 - 1}} \\ &= \sqrt{\frac{1539}{171}} \\ &= \sqrt{9} \\ &= 3 \end{aligned}$$

This formula is less reliable when the percentage in question (p) is very small or very large. If it is below 15% follow this procedure:

	Steps	Example (p = 10%, and n = 172)
1.	Calculate the standard error for p.	$SE = \sqrt{\frac{pq}{n - 1}} = \sqrt{\frac{10 \times 90}{171}} = 2.294$
2.	Double it	$2 \times 2.294 = 4.588$
3.	Add this to p and call it p' (pronounced as 'p prime')	$p' = 10 + 4.588 = 14.588$
4.	Calculate the standard error for p'.	$SE = \sqrt{\frac{p'q'}{n-1}} = \sqrt{\frac{14.588 \times 85.412}{171}}$ $= 2.7$

It is step 4, rather than step 1, that gives you the standard error. In this example, the standard error for your result of '10%' is 2.7%, not 2.3%. When p is greater than 85% you do the same thing except that step 3 would be:

> 'Take this away from p and call it p'.'

In the chapter on sampling, I promised to return to the question of how many people you should interview. In addition to practical considerations, such as the number of places you can visit in the time available, one point to think about is the level of accuracy that you require for your survey. For the kind of research I described in Chapter 1, results that are accurate to the nearest 10%, or even 15%, are usually adequate. If you were wondering how much emphasis to put on radio programmes for your correspondence students, and you conducted a survey to find out what proportion of them had access to a radio, a result of 'about 60% (almost certainly between 50% and 70%)' would probably be adequate. Whether the true figure was exactly 60% or, say, 66% would not matter.

If you can decide on the level of accuracy you require, you can use the formula for the standard error to calculate how large a sample you need. Make a guess at a probable result. Take this as p in the formula. Then calculate the 95% confidence limits that you would get using different sample sizes (i.e. different values of n).

For example, suppose you guessed that about 60% of all your students had access to a radio and you decided that you wanted an estimate accurate to the nearest 10%. The standard error for a result of '60%' would be 5.2 from a sample of 90, and 4.9 from a sample of 100. So a sample of between 90 and 100 would be sufficient.

A complication here is that you might be more interested in a part of the sample than in the whole sample. For example, having asked the students whether they had access to a radio, you might want to find out about the quality of reception from those who did have access to a radio. If you wanted your results from this subsample of students also to be accurate to the nearest 10%, your sample would have to contain about 90 to 100 students with access to radios. If you guessed that about 60% of the students would have access to radios, you would have to contact between 150 and 170 students altogether in order to get enough (i.e. 90 to 100) of those with radios.

The formula that I have been using so far is for a simple random sample or a systematic random sample. Most surveys, however, use a cluster sample and the formula for the standard error of a cluster sample is more complicated. (These types of sample are described on pages 41 - 44.) A shortened version of the formula, which is almost as good as the full one, is as follows:

$$SE = \sqrt{\left(\frac{\Sigma(\bar{p} - p)^2}{m - 1}\right)\left(\frac{1}{m}\right)}$$

p, as before, is the percentage in question, only now you calculate it separately for each cluster. $\bar{p}$ is the average of these percentages. m is the number of clusters.

Suppose you have divided all the villages in the country into six strata according to their size. You have drawn one village at random from each stratum and have interviewed 40 heads of households in each one. You have calculated, for each village, the proportion who own poultry, and the results are as follows:

40%, 50%, 50%, 55%, 65%, 70%

$\bar{p}$, the mean of these six percentages, is 55%.

First, you calculate $\Sigma(\bar{p} - p)^2$, as follows:

$\bar{p}$	p	$(\bar{p} - p)$	$(\bar{p} - p)^2$	
55	40	15	225	
55	50	5	25	
55	50	5	25	
55	55	0	0	
55	65	-10	100	←(Remember, -10 × -10 = 100,
55	70	-15	225	not -100.)

$$\Sigma(\bar{p} - p)^2 = 600$$

m, the number of villages, is 6, so (m - 1) = 5

$$\frac{\Sigma(\bar{p} - p)^2}{m - 1} = \frac{600}{5} = 120$$

$$\left(\frac{1}{m}\right) = \frac{1}{6}$$

$$\text{so, } \left(\frac{\Sigma(\bar{p} - p)^2}{m - 1}\right)\left(\frac{1}{m}\right) = 120 \times \frac{1}{6} = 20$$

Finally, you obtain the square root of this number:

$\sqrt{20} = 4.5$ (to one decimal place)

So, the standard error for this result ('55% of households own poultry') is 4.5%. The 95% confidence limits are 46% and 64%.

If you drew your cluster sample in a different way, or if you interviewed a different number of people in each village, the calculation of the standard error would get more complicated and you would need to consult a statistician.

A test of the difference between proportions

You can use this test if your results are in the form of a two-by-two table, like the results of the doctor's experiment that I used as an example in Chapter 11. The results were as follows:

	with drug %	without drug %
Survived	63	48
Died	37	52
	—	—
Base totals	(80)	(80)

The base totals in the two groups do not have to be the same, but the smaller of the two should not be less than 20. You can perform the test only if the results are from a random sample or if people are assigned to the two groups at random, as in the doctor's experiment. First you calculate the following:

$$2 \times \sqrt{\frac{p_1 q_1}{(n_1 - 1)} + \frac{p_2 q_2}{(n_2 - 1)}}$$

What this means is that you take the first group and you calculate

$$\frac{pq}{(n - 1)}$$

where p is the percentage in question, q is (100 - p), and n is the number of cases.

In the formula, this part appears as:

$$\frac{p_1 q_1}{(n_1 - 1)}$$

Then you do the same for the second group; this appears in the formula as:

$$\frac{p_2 q_2}{(n_2 - 1)}$$

Then you add these two amounts together, take the square root of the result, and finally multiply by 2.

The little numbers 1 and 2 in this formula are known as 'subscripts'; 'n_1' means the number of cases in the first group, 'n_2' the number of cases in the second group. Don't confuse a subscript 2, as in 'x_2', with the 2 that means 'squared' as in 'x^2'.

The calculation of this amount for the doctor's results would be as follows:

$$2 \times \sqrt{\frac{(63 \times 37)}{(80 - 1)} + \frac{(48 \times 52)}{(80 - 1)}}$$

$$= 2 \times \sqrt{\frac{2331}{79} + \frac{2496}{79}}$$

$$= 2 \times \sqrt{61.1}$$

$$= 2 \times 7.82$$

$$= 15.64$$

The larger the difference between the two groups (63% - 48%, in this example), the less likely it is to have occurred by chance. The test is whether the difference is larger than this figure you have calculated (15.64). If it is, the probability of it having occurred by chance is quite small - less than 5%. There is a difference of 15% between the doctor's two groups in the proportions who survived the disease (63% - 48%). Unfortunately for the doctor, this difference is not quite large enough to pass this test, since 15 is slightly less than 15.64. According to this test, the probability of the doctor's results occurring just by chance is slightly higher than 5%; in other words, his results fail to reach the 5% level of statistical significance. If the difference between the percent who survived in the two groups had been greater than 15.64 (say 70% had survived in one group and 48% in the other - a difference of 22%), this would have been statistically significant at the 5% level.

The χ^2 (chi-squared) test

The symbol χ is not a capital letter X, but the Greek letter chi (pronounced like the word 'kite' without the 't'). You do a χ^2 test in two stages. First you calculate something called 'χ^2' from your results; then you use this figure along with a set of tables to tell you the probability of getting your results by chance. If you are testing results from a survey, you can use this test only if the survey sample was drawn at random. If the results are from an experiment, you can use it only if people are assigned to the groups at random.

To calculate the value of χ^2 for your results, you use the following formula:

$$\chi^2 = \Sigma\frac{(O - E)^2}{E}$$

In this formula, O (the letter 'O') stands for 'Observed result' and E stands for 'Expected result', and I'll use the doctor's results again to explain what these terms mean.

The 'observed results' are simply the results that the doctor actually got. You could present these results with the column totals and the row totals filled in, like this:

	Patients with drug	Patients without drug	Totals
Survived	50	38	88
Died	30	42	72
Totals	80	80	160

The 'expected results' are the results that you would expect to get if the drug had absolutely no effect. Suppose you had a table of results with the totals the same as in the above one, but with the observed results removed, like this:

	Patients with drug	Patients without drug	Totals
Survived			88
Died			72
Totals	80	80	160

Out of the grand total of 160 patients, 88 survived, making a survival rate of 55%. If the drug had no effect, there would be no reason to expect any difference between the two groups; in other words, you would expect the same survival rate in each of the two groups - 55% in the group with the drug and 55% also in the group without the drug. 55% of 80 is 44, so you would expect 44 survivors in each group. These are the 'expected results', i.e. the results that you would expect if the drug had no effect. You can put them in the table as follows:

	Patients with drug	Patients without drug	Totals
Survived	44	44	88
Died	36	36	72
Totals	80	80	160

The following diagram may help you to remember how to calculate 'expected results':

	Column 1	Column 2	Totals
Row 1	A	B	S
Row 2	C	D	T
Totals	U	V	W

The expected value for A is $U \times \frac{S}{W}$

The expected value for B is $V \times \frac{S}{W}$

The expected value for C is $U - A$

The expected value for D is $V - B$

Always check that the expected results you have calculated add up correctly to the row and column totals; A and B, for example should add up to S, while A and C should add up to U.

Alongside each of the observed results in the doctor's table, we can now put the expected results, in brackets. Or, saying the same thing in different words, in each of the four cells of the table we can put the observed and expected results, as follows:

	Patients with drug	Patients without drug
Survived	50 (44)	38 (44)
Died	30 (36)	42 (36)

The instructions contained in the formula for χ^2 are to be read as follows:

1 Subtract the expected result from the observed result

2 Square what you get

$$\Sigma \frac{(O - E)^2}{E}$$

4 Do this for each cell and add up what you've got

3 Divide this by the expected result

For the doctor's results, the calculations are as follows:

	O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Cell A	50	44	6	36	0.818
Cell B	38	44	-6	36	0.818
Cell C	30	36	-6	36	1.0
Cell D	42	36	6	36	1.0

$$\Sigma\frac{(O - E)^2}{E} = 3.636$$

The value of χ^2, for the doctor's table of results, is 3.636. The next step would be to consult a table of values for χ^2 in a book of statistical tables, but I will defer my explanation of that stage until I have given another example. Basically, the higher the value of χ^2, the less likely it is that the results occurred by chance. What the published tables would tell you, in this case, is that if the value of χ^2 that you have calculated is over 3.84, there is less than a 5% probability of getting those results just by chance. The value of χ^2 for the doctor's results is 3.636, just slightly short of 3.84, so the verdict of the test is that the doctor's results are not statistically significant at the 5% level.

The tables would also show, incidentally, that if the value of χ^2 is over 2.71, there is less than a 10% probability of getting these results by chance; so, though the doctor's results are not quite significant at the 5% level, they are significant at the 10% level. Note that the higher the value of χ^2, the lower the probability of getting those results by chance.

The χ^2 value is a statistic which shows how far the observed results in a table depart from the expected results. The greater the difference, the larger the value of χ^2 and the lower the probability of those results occurring by chance. On page 148 I gave two alternative sets of results that the doctor might have got. If you calculate the χ^2 values for these tables (and you might like to do so to check that you have followed the procedure correctly), you should get 33.629 for Table B and 0.025 for Table C. The results of Table B, most unlikely to occur by chance (the probability is well below one in 10,000), give a very high value for χ^2; those of Table C give a very low value.

In these examples, it is the size of the difference between the two groups that is the important thing; the difference between the doctor's two groups is not statistically significant at the 5% level, whereas the much larger difference in Table B is highly significant. The size of the sample is also important. Suppose that the doctor had got the same result (in terms of percentages), but that he had used ten times as many patients, i.e. 1,600 instead of 160. His results would have been as follows:

	Patients with drug	Patients without drug
Survived	500	380
Died	300	420

The χ^2 value for these results is 36.364. Multiplying the sample by ten, while keeping the proportions the same, multiplies the χ^2 value by ten. These results would be highly significant, demonstrating the effectiveness of the drug beyond all doubt. Because χ^2 reflects the size of the sample as well as the size of the difference, you always calculate χ^2 from the actual results, never from percentages.

The table of the doctor's results has just two rows and two columns. Tables of this size are called 'two-by-two' tables. A table with, say, four rows and five columns would be a 'four-by-five' table. The χ^2 test can be used with tables of any size. Say you obtained the following two-by-three table of results from a survey of radio-listening:

	Educational level		
	No schooling	Reached St.1-3	Reached St.4 or over
	%	%	%
Listened at least once in previous week	14	22	32
Did not listen at all in previous week	86	78	68
Base totals	(110)	(76)	(50)

It appears that there is a clear relationship between educational levels and radio listening; the more educated seem to listen to the radio more often. But you should perform a test of statistical significance on the results, just to make sure that there is no serious risk of these results having occurred by chance. First, you take the raw figures:

	Educational level			
	No schooling	Reached St. 1-3	Reached St. 4 or over	Totals
Listened at least once in previous week	15	17	16	48
Did not listen at all in the previous week	95	59	34	188
Totals	110	76	50	236

Using the diagram on page 253, you can see that the expected result for the first cell - the one to go alongside the observed result of 15 - is 110 × $\frac{48}{236}$, which comes to 22.4. Similarly, the expected result for the second cell is 76 × $\frac{48}{236}$, which comes to 15.5. The table of observed and expected results and then the calculation of X^2 is as follows:

15 (22.4)	17 (15.5)	16 (10.1)
95 (87.6)	59 (60.5)	34 (39.9)

O	E	O-E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
15	22.4	-7.4	54.76	2.445
17	15.5	1.5	2.25	0.145
16	10.1	5.9	34.81	3.447
95	87.6	7.4	54.76	0.625
59	60.5	-1.5	2.25	0.037
34	39.9	-5.9	34.81	0.872
				7.571

The χ^2 value for this table is 7.571. The critical value of 3.84 applies only to two-by-two tables. This is a two-by-three table. To find out what a χ^2 value of 7.571 tells you for a two-by-three table, you have to consult a set of χ^2 tables (see Appendix 7, page 308). A part of the table will look like this:

	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28

Across the top you will find the levels of significance. The column headed '5%' is usually the one you are interested in, though you might sometimes want the ones headed 10% and 1%. Down the left-hand side you will find a set of numbers called 'degrees of freedom'.

Every table of results has a certain number of degrees of freedom. You calculate this by multiplying the number of rows minus one by the number of columns minus one. For a two-by-three table, you calculate $(2 - 1) \times (3 - 1)$. This, of course, is $1 \times 2 = 2$. So a two-by-three table has two degrees of freedom. A two-by-two table would have one degree of freedom, since $(2 - 1) \times (2 - 1) = 1$. A four-by-five table would have $(4 - 1) \times (5 - 1) = 3 \times 4 = 12$ degrees of freedom.*

For a table with two degrees of freedom, such as the two-by-three table in this example, you consult the second line of the χ^2 tables. There you will find that the value of χ^2 at the 5% level is 5.99. 7.571 is larger than 5.99, so the differences shown in the table on radio listening are statistically significant at the 5% level. It is unlikely that you got these results just by chance, so you can conclude with confidence that people who have had more schooling listen to the radio more often.

---

* It is not essential to know what the expression 'degrees of freedom' means. To give you an idea, however, look again at the two-by-two table of the doctor's results on page 253 with the row and column totals filled in but the actual results removed. If you put a figure into any one of the four empty cells, you can then complete the remaining three cells. Suppose all you were told was that 20 patients with the drug died. You enter this figure in the table. You can then fill in the other three since it follows that 60 patients with the drug survived (80 - 20), 52 patients without the drug died (72 - 20) and 28 patients without the drug survived (88 - 60, or 80 - 52). You might say that just one cell in this table (any one of the four) is free to take any value and that the other three cells are then fixed. So a two-by-two table has just one degree of freedom. Don't worry if you don't understand this. You only need to know how to calculate the number of degrees of freedom; you don't need to understand what they are.

It is generally best to exclude 'Inadequate information' respondents before doing a χ^2 test on a table. Take the following example:

	Lowlands	Mountains
Had read a newspaper in the previous week	24	9
Had not read a newspaper in the previous week	54	20
Inadequate information	2	11

If you performed a χ^2 test on these figures, you would find that the difference was statistically significant, even at the 1% level. But this is entirely due to the difference in the numbers coded as 'Inadequate information', i.e. the test would show that a significantly higher proportion of mountain people had been coded as 'Inadequate information'. The test would tell you nothing at all about differences in newspaper reading. In fact, if you disregard the 'Inadequate information' category, there is no difference at all between the lowlands people and the mountain people in their newspaper reading; 31% had read a newspaper in the previous week, for both groups. Incidentally, this means that, if you get a computer to calculate χ^2 values for the tables that it produces for you, you must check to see what tables the computer actually used in its calculations. Unless specifically instructed not to do so, the computer will include all categories in the table (Inadequate information, Don't know, Does not apply, and so on), so the χ^2 values might be quite misleading.

The χ^2 test is less reliable if any of the numbers in the table are very small. There are two things you can do about this. One is to rearrange the table so as to remove the very small numbers. Say you had the following raw figures:

	Correspondence students studying		
	Maths.	Biology	Agricultural science
Have completed first worksheet	28	16	1
Have not completed first worksheet	47	53	7

Before doing a χ^2 test on this table, you should either exclude the Agricultural science students completely, or you could add them to the Biology students to produce the following table:

	Correspondence students studying Maths.	Sciences (Biol. and Ag. Sci.)
Have completed first worksheet	28	17
Have not completed first worksheet	47	60

If it is not possible to rearrange the table, you can make what is called a 'continuity correction'. You should do this if any one of the expected results is less than 1, or if more than a fifth of the expected results in the table are less than 5. Take the following example:

	Correspondence students using Printed course only	Printed course and radio programmes
Have completed first worksheet	2	9
Have not completed first worksheet	17	14

The expected result for the first cell is 4.8. There are only four results in a two-by-two table. One out of four is one quarter, which is more than one fifth, so you should apply the continuity correction. This means that you calculate O - E as usual (2 - 4.8 = - 2.8), but then you reduce this figure by taking off 0.5 before squaring it, i.e. you do not calculate $(-2.8)^2$, but instead $(-2.3)^2$. The χ^2 value for this table, calculated without the continuity correction, would be 3.917, i.e. greater than 3.84 and therefore statistically significant for a two-by-two table at the 5% level. If you make the continuity correction, however, the value is only 3.386, i.e. not statistically significant at the 5% level.

Standard deviation

To calculate the standard deviation of a set of values, such as test scores, you first calculate the mean. Then follow these steps:

1. Take each score and subtract the mean.
2. Square the result.
3. Add up these amounts.
4. Divide by the number of scores.
5. Take the square-root.

As a formula, it looks like this:

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$$

Here are the calculations for the standard deviation of a set of scores (those of group A on page 154). (Remember that a minus number multiplied by another minus number makes a plus number. For example, 48 - 50 = -2, and - 2 × - 2 = 4, not -4.)

Score (x)	Score minus the mean $(x - \bar{x})$	Squared $(x - \bar{x})^2$
62	12	144
60	10	100
56	6	36
54	4	16
48	-2	4
45	-5	25
43	-7	49
42	-8	64
40	-10	100
	Total $\Sigma(x-\bar{x})^2 =$	538

$$\frac{\Sigma(x - \bar{x})^2}{n} = \frac{538}{9} = 59.778$$

$$\sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{59.778} = 7.73 \qquad \text{(to two decimal places)}$$

The standard deviation of group B's scores, calculated in the same way, is 39.03 - a higher figure, indicating that there is more variation in the scores of group B.

An alternative formula, which gives the same result but makes the calculations easier, is as follows:

$$s = \sqrt{\frac{n\Sigma x^2 - (\Sigma x)^2}{n^2}}$$

Dependent and independent scores

Before describing any significance tests for use with measurement variables, I have to introduce yet another complication. You use different tests depending on whether your sets of scores are independent of each other or not.

Suppose a researcher has divided 60 students into two groups at random and given them all a test. Suppose you were given the names of two students in group B - Beth Brown and Belinda Barrymore - and you were asked to guess which had done better. Would it help you if you were given full details of the scores obtained by group A (but none for group B)? Clearly not. You might see from the scores of group A that, say, Ann Ames had done better than Angela Appleyard but how would this help you make a guess about Beth Brown and Belinda Barrymore? It wouldn't. The two sets of scores are independent of each other.

But suppose that the researcher had arranged the students in matched pairs; he had put them in pairs according to their results on a recent test, and then assigned one member of each pair to group A, the other to group B, at random by tossing a coin. If you were given the scores of group A and also told which members of group B were the 'partners' of those in group A, this would help you to make your guess. If you knew that Beth had been paired with Ann and Belinda with Angela, and that Ann had done better than Angela, you would guess that Beth had done better than Belinda. As a result of the matching, you expect some relationship between the scores of group A and those of group B. The two sets of scores are not completely independent of each other, or, to put it the other way round, they are dependent.

Another common example of dependent sets of scores is when you give a pre-test and a post-test to the same group of people. If you were asked to guess whether Tom did better than George on the post-test, would it help you if you were told that Tom did better than George on the pre-test? Of course it would. The rank-order of the students will probably be similar on both tests, with the better students getting higher scores on both. There will be a clear relationship, or dependence, between the two sets of scores.

If you are doing these tests on scores from two sets of students - and this includes both matched and unmatched groups - it is essential that the students were assigned to the groups at random; otherwise the tests do not apply.

For independent sets of scores:

If you have at least 15 students in each group, if the groups are roughly the same size (i.e. the smaller is at least three-quarters the size of the larger), if there is a similar amount of variation within the scores of each group and if the shape of the distributions is not markedly out of the ordinary, use the t test for independent sets of scores (page 261). Otherwise use the Mann-Whitney U test (page 264).

For dependent sets of scores:

If you have at least 15 pairs of scores, if there is a similar amount of variation in the two sets and if the shape of the distribution is not markedly out of the ordinary, use the t test for dependent sets of scores (page 267). Otherwise use the Wilcoxon matched-pairs signed-ranks test (page 269).

The 't test' for independent sets of scores

Suppose a researcher has given two versions of a lesson to two groups of students and then given them all the same test at the end. Group 2 did slightly better than group 1, on average; can he conclude that group 2's version of the lesson was more effective? In other words, is the difference statistically significant?

He assigned students to the two groups at random, without matching, so the two sets of scores are independent. Let's assume that he has drawn diagrams of the two distributions and found that they are not out of the ordinary, so he can proceed with a significance test.

There are many tests he can use, one of which is called the 't test'. As with many significance tests, you first calculate a figure from your results - this time it's something called 't' - and then you look it up in published tables to find the probability of getting such a figure purely by chance. As I explained in Chapter 11, the statistical significance is affected by the amount of variation within the groups, as well as by the size of the difference between them. Consequently, the calculations are somewhat complicated. The formula for t, which looks a bit fierce, is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}\right)}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}$$

There are several steps in calculating t, and that's why the formula is complex, but each individual step is quite easy (with a calculator). If you go through them systematically, it's not too difficult. The steps are as follows:

1. Calculate the mean of group 1. This is $\bar{x}_1$.

2. Subtract each score in group 1 from the means and square what you get. Each of these is $(x_1 - \bar{x}_1)^2$.

3. Add them up. This is $\Sigma(x_1 - \bar{x}_1)^2$. It's the sum of the squared deviations for group 1.

4. Do the same for group 2. This is $\Sigma(x - \bar{x}_2)^2$.

5. Add together the sums of squared deviations for groups 1 and 2. This is $\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2$.

6. Add together the number of cases in group 1 (n_1) and the number in group 2 (n_2) and then subtract 2. This is $n_1 + n_2 - 2$.

7. Divide the result of step 5 by the result of step 6 and take the square root. This is

$$\sqrt{\left(\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}\right)}$$

8. Add n_1 and n_2.

9. Multiply n_1 by n_2.

10. Divide the result of step 8 by the result of step 9 and take the square root. This is

$$\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

11. Multiply the result of step 7 by the result of step 10.

12. Subtract the mean of group 2 from the mean of group 1. This is $\bar{x}_1 - \bar{x}_2$.

13. Divide the result of step 12 by the result of step 11, and that's the value of t.

Normally, for an experiment of this kind (comparing two versions of a lesson) you should have at least 15 students in each group and peferably more. However, to save space, I'll calculate t for results from an experiment in which group 1 consisted of seven students and group 2 of nine. (You don't have to have the same number in both, though it's better to have roughly the same.) Each score is out of ten:

Scores of group 1

	(x_1)	$(x_1 - \bar{x}_1)$	$(x_1 - \bar{x}_1)^2$
	2	-3	9
	3	-2	4
	4	-1	1
	5	0	0
	5	0	0
	7	2	4
	9	4	16
Totals	35		34

Scores of group 2

	(x_2)	$(x_2 - \bar{x}_2)$	$(x_2 - \bar{x}_2)^2$
	3	-4	16
	5	-2	4
	6	-1	1
	7	0	0
	7	0	0
	8	1	1
	8	1	1
	9	2	4
	10	3	9
Totals	63		36

$\bar{x}_1 = 35 \div 7 = 5$

$\bar{x}_2 = 63 \div 9 = 7$

$$t = \frac{5 - 7}{\sqrt{\frac{34 + 36}{7 + 9 - 2}}\ \sqrt{\frac{7 + 9}{7 \times 9}}}$$

$$= \frac{-2}{\sqrt{\frac{70}{14}}\sqrt{\frac{16}{63}}} = \frac{-2}{\sqrt{5}\sqrt{0.254}}$$

$$= \frac{-2}{2.236 \times 0.504} = \frac{-2}{1.127}$$

$$= -1.775$$

Intuitively you'd expect a larger difference between the means to be more significant (i.e. less likely to be due to pure chance). You'd also expect that a lot of variation in the scores within the groups would make the result less significant. And you'd also expect a given result to be more significant if it was based on larger groups of students. The formula for t takes these three things into account.

A big difference between the means makes this a larger number, which makes t larger.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}\right)}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}$$

A lot of variation makes this a large number, which makes t smaller.

Larger groups make this smaller, which makes all the bottom part smaller, which makes t larger.

The larger the value of t, the more statistically significant the result. As for chi-squared, tables of t are set out with various levels of significance along the top and degrees of freedom down the side. The degrees of freedom are the total number of students minus two ($n_1 + n_2 - 2$). For the example I have just calculated, there are (7 + 9 - 2) = 14 degrees of freedom. The t tables (see Appendix 7, page 309) show that, with 14 degrees of freedom, a value of 1.775 is significant at the 10% level (it's greater than 1.761) but not at the 5% level (it's less than 2.145). You can ignore the fact that t has a minus value in this example. That simply indicates that $\bar{x}_2$ was higher than $\bar{x}_1$ (i.e. group 2 did better). It's the difference between the two groups that you're testing, regardless of which did better.

The Mann-Whitney U test

The t test that I've just described makes assumptions about certain features - or 'parameters' - of the distributions. Tests that do this are sometimes called 'parametric tests'. There are other tests which do not make these assumptions, so they are called 'non-parametric tests'. The Mann-Whitney U test is one of these. It disregards the actual scores that the students got and concentrates only on the rank order of their scores.

You can use the Mann-Whitney U test when one of your groups is much smaller than the other. The larger of the two groups should contain at least nine members and preferably more. As always, a given result is more statistically significant when based on larger groups, so if you want to be confident about your findings, it's better to use larger groups. This test, as compared with the t test, is less likely to be upset if one group has more variation than the other or if the shape of the distributions is slightly odd. However, as I said in Chapter 11, if either of the distributions has a markedly odd shape, you ought to find the reason for the odd shape rather than just proceed with a significance test.

You begin by putting the scores of all the students (i.e. both groups combined) into order from highest to lowest. Suppose group 1 consisted of ten students and group 2 of twelve, and they got the following scores:

Group 1 75, 72, 70, 64, 61, 55, 53, 50, 45, 36

Group 2 89, 85, 80, 79, 74, 70, 68, 65, 62, 59, 47, 40

You put all 22 scores in order, as in the column on the left:

	Rank	
89	1	
85	2	
80	3	
79	4	
75	5	1
74	6	
72	7	1
70	8.5	
70	8.5	1
68	10	
65	11	
64	12	1
62	13	
61	14	1
59	15	
55	16	1
53	17	1
50	18	1
47	19	
45	20	1
40	21	
36	22	1

Next you write the rank order of each score next to it, starting with number 1 for the highest. The only problem here is if you have two or more scores the same. In this example, the eighth and ninth students had both scored 70. You give the same average rank to both of them. The average of 8 and 9 is 8.5, so they are both ranked as 8.5. If the eighth, ninth and tenth students had got the same score, they would all have been ranked as 9, since 9 is the average of 8, 9 and 10 (27 ÷ 3 = 9), and the next student in the list would have been ranked 11.

Next you mark those that belong to group 1, and then you add up the ranks of group 1, thus:

Group 1's ranks = 5 + 7 + 8.5 + 12 + 14 + 16 + 17 + 18 + 20 + 22 = 139.5

Call this number R_1 and use it to calculate something called 'U' for group 1, with the following formula (n_1 as before, is the number of students in group 1, and n_2 the number in group 2):

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

$$\begin{aligned} U_1 &= (10 \times 12) + (\frac{10 \times 11}{2}) - 139.5 \\ &= 120 + 55 - 139.5 \\ &= 35.5 \end{aligned}$$

Now use this number to calculate U for group 2, by this formula:

$$U_2 = n_1 n_2 - U_1$$

$$\begin{aligned} U_2 &= (10 \times 12) - 35.5 \\ &= 84.5 \end{aligned}$$

Compare U_1 and U_2 and take the smaller of the two. This is your value for U.

You compare this figure with published tables. The procedure is slightly different from the one you use for t and chi-squared, since there are several sets of tables, one for each level of significance (see Appendix 7, page 310). Choosing, say, the 5% level, you look up the value of U for when one group has ten members and the other has twelve (i.e. $n_1 = 10$, $n_2 = 12$). You'll find that this value is 29. The difference between the two groups is significant at the 5% level if your value of U is smaller than the one in the table. In my example, 35.5 is larger than 29, so the difference between the two groups is not significant at the 5% level. Turning to the table for the 10% level, you'll find that the value of U when one group has ten members and the other has twelve is 34. 35.5 is larger than 34, so the difference between these groups is not quite significant at the 10% level.

To get an idea of how U reflects the amount of difference between the two groups, consider two groups of the same size. If there is a similar range of scores in each group, the students of group 1 will be spread evenly over the rank order list - some at the top end, some at the bottom and some in the middle. The total of ranks for group 1 will be much the same as for group 2, so U_1 will be similar to U_2, and neither of them will be very small. By contrast, if group 2's students nearly all did better than group 1's, all the ranks at the top end (1, 2, 3 etc.) would go to group 2's students and all those at the bottom end to group 1's. As a result, the total of ranks (and therefore the value of U) would be very large for group 1 and very small for group 2. The smallness of the smaller value (group 2's) would indicate a big difference between the groups.

The tables for U in Appendix 7 only give values for when there are up to 20 students in each group. If you have more than 20 students in either group, you use your value U to calculate yet another statistic, this time called z. The formula is as follows:

$$z = \frac{U - \left(\frac{n_1 n_2}{2}\right)}{\sqrt{n_1 n_2 \frac{(n_1 + n_2 + 1)}{12}}}$$

Say you had 22 people in group 1, 25 in group 2 and a value of U of 165, you would calculate z as follows:

$$z = \frac{165 - \frac{(22 \times 25)}{2}}{\sqrt{22 \times 25 \times \frac{(22 + 25 + 1)}{12}}}$$

$$= \frac{165 - 275}{\sqrt{550 \times \frac{48}{12}}} = \frac{-110}{\sqrt{550 \times 4}}$$

$$= \frac{-110}{\sqrt{2200}} = \frac{-110}{46.904}$$

$$= -2.345$$

You can get the significance level from z very easily with the following table:

Value of z	Significance level
1.29	20%
1.44	15%
1.68	10%
1.96	5%
2.58	1%

The higher the value of z (regardless of whether it is positive or negative), the more statistically significant the result. A value of 2.345 for z, being larger than 1.96, is significant at the 5% level.

The t test for dependent sets of scores

You can use t for dependent sets of scores as well as independent ones, but you calculate it differently. I'll give the example of an experiment in which fifteen students did a pre-test, then read a correspondence lesson, and then did the same test again as a post-test. The procedure would be exactly the same if, say, you had arranged thirty students into pairs of equal ability, then divided each pair at random between the two groups and given a different version of the lesson to each group. You begin by calculating the difference between each pair of scores, as in the table which follows; the rest of the test concentrates solely on these differences. Each score is out of 20.

Pre-test scores (x_1)	Post-test scores (x_2)	Difference, called d ($x_2 - x_1$)	($d - \bar{d}$)	$(d - \bar{d})^2$
2	4	2	0.4	0.16
4	4	0	-1.6	2.56
5	6	1	0.6	0.36
5	4	-1	-2.6	6.76
7	8	1	-0.6	0.36
9	11	2	0.4	0.16
9	13	4	2.4	5.76
10	11	1	-0.6	0.36
10	10	0	-1.6	2.56
10	12	2	0.4	0.16
11	13	2	0.4	0.16
12	15	3	1.4	1.96
14	18	4	2.4	5.76
15	19	4	2.4	5.76
17	16	-1	-2.6	6.76
Totals		24		39.6

$$\bar{d} = \frac{24}{15} = 1.6 \qquad S_d = \sqrt{\frac{39.6}{15}} = \sqrt{2.64} = 1.625$$

The first two columns give the two sets of scores and the third gives the difference between them, e.g. the first student scored 2 on the pre-test and 4 on the post-test, a difference of 2 marks. Note that the fourth student actually did worse on the post-test, so the difference for him is - 1, not 1. You add up the positive differences and subtract the negative ones (the minus signs are important here) to get the total of the differences (24). Dividing by the number of pairs (15) gives the mean difference, which I've called $\bar{d}$. In this example, it comes to 1.6.

Next you calculate the standard deviation of the differences (see page 259). For the first one, you calculate 2 - 1.6 = 0.4 and then you square it: 0.4 × 0.4 = 0.16. You repeat this for each difference; note that -1 - 1.6 = -2.6. Then you add up the results (39.6), divide by the number of differences (15) and take the square root. This gives you the standard deviation of the differences, which I've labelled S_d. In the example, it's 1.625.

Using the results of these calculations, you compute t according to this formula:

$$t = \frac{\bar{d}}{\left(\frac{S_d}{\sqrt{n - 1}}\right)}$$

$$t = \frac{1.6}{\left(\frac{1.625}{\sqrt{15 - 1}}\right)} = \frac{1.6}{\left(\frac{1.625}{\sqrt{14}}\right)}$$

$$= \frac{1.6}{\left(\frac{1.625}{3.742}\right)} = \frac{1.6}{0.434}$$

$$= 3.685$$

You compare this figure with the values of t shown in the t tables (Appendix 7, page 309). The degrees of freedom, for results in pairs like these, are the number of pairs minus one: in this example 15 - 1 = 14. From the tables you can see that, with 14 degrees of freedom, a value of 3.685 for t is significant at the 5% level (it's bigger than 2.145), and even at the 1% level (it's bigger than 2.977). If you look back at the results, this isn't really surprising. All except four of the students did better on the post-test and only two actually did worse. Such a consistent improvement throughout the group is most unlikely to be due to chance.

The Wilcoxon Matched-pairs Signed-ranks test

Suppose you arranged thirty students into pairs of roughly equal ability and decided by tossing a coin which member of each pair should go into group 1 and which into group 2. They have studied different versions of a lesson and taken the same test at the end. (You would use the same test if the scores were pre-test and post-test results for a single group of fifteen students.) As with the t test for paired scores, you begin by calculating the difference between each pair. For the moment, you can ignore whether the group 2 score was higher or lower than the group 1 score. Each score is out of 100.

Group 1's scores	Group 2's scores	Difference	Rank of difference	Sign of difference
32	42	10	10	+
37	34	3	1	-
40	40	0	none	neither
45	33	12	11	-
52	67	15	13	+
59	64	5	3	+
60	74	14	12	+
62	68	6	5	+
67	84	17	14	+
70	79	9	9	+
73	68	5	3	-
79	72	7	6	-
81	86	5	3	+
85	93	8	7.5	+
89	97	8	7.5	+

If any pair got the same score, like the third pair in this example, you ignore them from now on. You rank the differences in order, giving rank 1 to the smallest difference. In the example, the second difference (3) is the smallest, so this is ranked 1. Then there are three differences of 5, so you give each of these the average of the three ranks that they share; they share the ranks 2, 3 and 4, and the average of these ranks is 3, so each gets ranked 3. The next largest is ranked 5, and so on. The two differences of 8 share ranks 7 and 8, so each is ranked 7.5.

Next you write the sign of each difference, i.e. whether the group 2 score was higher than the group 1 score (+) or lower (-). In the example there were ten + signs and four - signs (making a total of 14, not 15, because we are leaving out the third pair).

Then you add up the total ranks of those marked '-': 1 + 11 + 3 + 6 = 21. And you do the same for those marked '+': 10 + 13 + 3 + 12 + 5 + 14 + 9 + 3 + 7.5 + 7.5 = 84. You compare these figures and take the smaller of the two. It's called T (capital T not to be confused with small t). For this set of results, the number of pairs is 14 (remembering that we're leaving one out) and T equals 21.

You compare this figure with published tables of T (Appendix 7, page 311). Your results are statistically significant if your value of T is equal to or less than the one shown in the tables. For 14 pairs, the value of T at the 5% level is 21. The T value for our results was 21, so the results are statistically significant (only just) at the 5% level.

You can see how the test works if you think of a set of scores in which about half of group 2 did better than their partners in group 1, and half did worse, and by similar amounts. There would be roughly equal numbers of + and - signs and the total of '+' ranks would be close to the total of '-' ranks; neither total would be very small. If, however, most of group 2 did better than their partners in group 1, and the few who did worse did not do a lot worse, there would be more + signs than - signs and the total of '+' ranks would be much bigger than the total of '-' ranks. The smallness of the smaller total (in this case the '-' ranks total) would indicate that there was a large difference between the groups.

The tables give values of T for up to 25 pairs. If you have more than 25 pairs, you calculate T as above, and then use this value to calculate z, according to the following formula:

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

N, in this formula, is the number of pairs. Suppose you had calculated a T of 110 from 28 pairs:

$$z = \frac{110 - \frac{28 \times 29}{4}}{\sqrt{\frac{28 \times 29 \times 57}{24}}}$$

$$= \frac{110 - 203}{\sqrt{1928.5}} = \frac{-93}{43.915}$$

$$= -2.118$$

From the table I gave for z on page 267, you can see that a value of 2.118 for z is statistically significant at the 5% level (it's bigger than 1.96).

Appendix 2 Statistical theory

It is possible to drive a car competently and safely - and many people do - without having any idea how it works. Something similar can be said about statistics; you can use a significance test, such as those in Appendix 1, without understanding how the calculations lead to the conclusions. I have to admit that I myself have only a vague notion of what the chi-squared statistic actually is and of how tables of chi-squared values are calculated. However, in doing statistics as in driving a car, it helps if you have some understanding of what's going on. This appendix describes briefly some important concepts underlying the calculation of confidence limits and statistical significance.

Imagine you have a large sack full of beads, each about the size of a small pea. Half the beads are red and half are blue, and they are thoroughly mixed up. You put your hand in the sack with your eyes closed, and take out four beads, one at a time. This forms a random sample of four taken from a large population (there are thousands of beads in the sack). As you take each one out, you record whether it's red or blue. There are sixteen different ways in which your sampling can proceed, and we can list them all, as follows:

	First bead	Second bead	Third bead	Fourth bead	Proportion of red beads
1	red	red	red	red	ALL
2	red	red	red	blue	3/4
3	red	red	blue	red	3/4
4	red	red	blue	blue	2/4
5	red	blue	red	red	3/4
6	red	blue	red	blue	2/4
7	red	blue	blue	red	2/4
8	red	blue	blue	blue	1/4
9	blue	red	red	red	3/4
10	blue	red	red	blue	2/4
11	blue	red	blue	red	2/4
12	blue	red	blue	blue	1/4
13	blue	blue	red	red	2/4
14	blue	blue	red	blue	1/4
15	blue	blue	blue	red	1/4
16	blue	blue	blue	blue	NONE

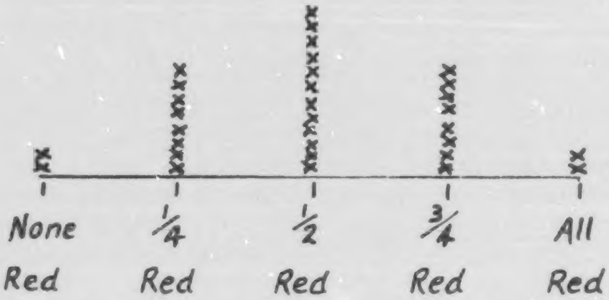
Every time you put your hand in, you have an equal chance of getting a red one or a blue one, because half the beads in the sack are red and half are blue. It follows that each of these sixteen possible samples is equally likely; there is one chance in sixteen (6.25%) of getting red-red-red-red, one chance in sixteen of getting red-red-red-blue, and so on. Putting it another way, if you repeated this procedure over and over again, throwing each sample back in the sack, jumbling them up and then taking a fresh sample, about one sixteenth of your samples would be red-red-red-red, one sixteenth would be red-red-red-blue, and so on.

One of the possible samples contains all red, and one contains all blue; the other fourteen contain a mixture. Looking at them in terms of the proportion of red beads in the sample, we can tabulate them as follows:

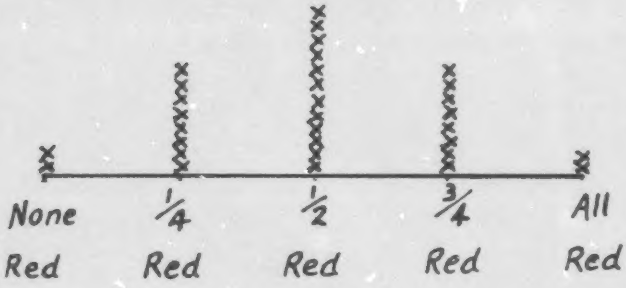
Proportion of red beads in the sample	Number of samples	%
None red	1	6.25
One-quarter red	4	25
Half red	6	37.5
Three-quarters red	4	25
All red	1	6.25
total	16	100

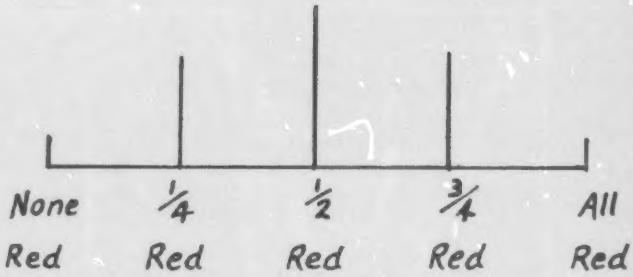
If you did this sampling many times, you'd find that about 6.25% of your samples gave a result 'None red', about 25% gave a result 'One quarter red', and so on. A hundred samples would give results quite close to these; a thousand samples would be even closer, and so on. The more times you did it, the closer you'd get to these figures.

Let's suppose that you did it 32 times and that, as it happened, you got exactly these proportions. You could, if you wanted, draw a diagram of your results, marking each sample result with a cross:



This diagram resembles those I gave in Chapter 11. Like them, it presents a distribution. But there is an important difference. In the earlier diagrams, each circle represented one student's test score. In this diagram, each cross represents a result derived from one sample; it presents the results of 32 separate surveys, so to speak, each of which took a sample of four beads. It is not a distribution of people; it is a distribution of results from random samples. What's more, it is a theoretical distribution. It doesn't show what someone really got from doing this sampling 32 times. In fact, if someone really went through this procedure, his pattern of results would probably not be exactly like this, though it would resemble it. It shows the pattern of results that you'd tend to get if you went on doing this sampling for ever. In order to show that the diagram presents a theoretical distribution of imaginary results, not a distribution of real results, I'll replace the columns of crosses with straight lines. The height of a line represents the relative frequency of that type of sample; for instance, the centre line, representing samples with half red, is six times higher than the line representing samples with none red:



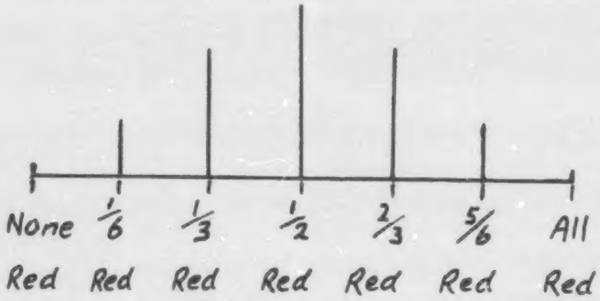


Now, what distribution would you get if you took samples of six beads each time, rather than four? I could list all the possible sequences of six beads, as I did for samples of four beads, but, with samples of six beads, there are 64 possible sequences and it would be tedious to write them all out. Take it from me that you get this distribution:

	Number of samples	%
None red	1	1.56
One sixth red	6	9.38
One third red	15	23.44
Half red	20	31.25
Two thirds red	15	23.44
Five sixths red	6	9.38
All red	1	1.56
Total	64	100

With samples of six beads each time, you expect to get samples containing no red beads once every 64 samples, on average; samples containing one red bead 6 times out of 64, on average, and so on.

In diagram form, it looks like this:

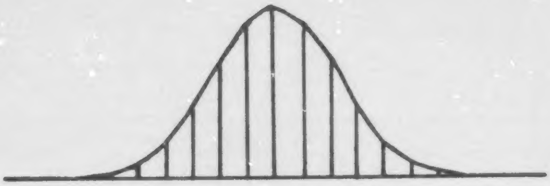


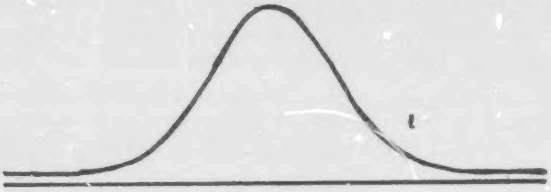
On the next page, I repeat the diagrams I've already given, showing the theoretical distribution of results that you'd get from drawing samples of four beads and then of six beads. Then I give, in the same form, diagrams to show the distributions of results you'd get from drawing samples of eight, ten, twelve, sixteen and twenty.

With diagrams on this scale, some of the vertical lines are too small for me to draw. For instance, with samples of 20 beads, it is possible to get a sample with no red ones, but this would happen very rarely - on average once every 1 048 576 samples, to be precise. This would be represented by a tiny vertical line at the left hand end of the bottom diagram; in relation to the centre line, which I've drawn as 25 mm high, it would be 0.00014 mm. You should think of all these distributions as tailing off right to the ends. Even if you were taking samples of 100 or 1 000 beads each time, there would be a possibility - an extremely small one - of getting a sample consisting entirely of red beads or entirely of blue ones.

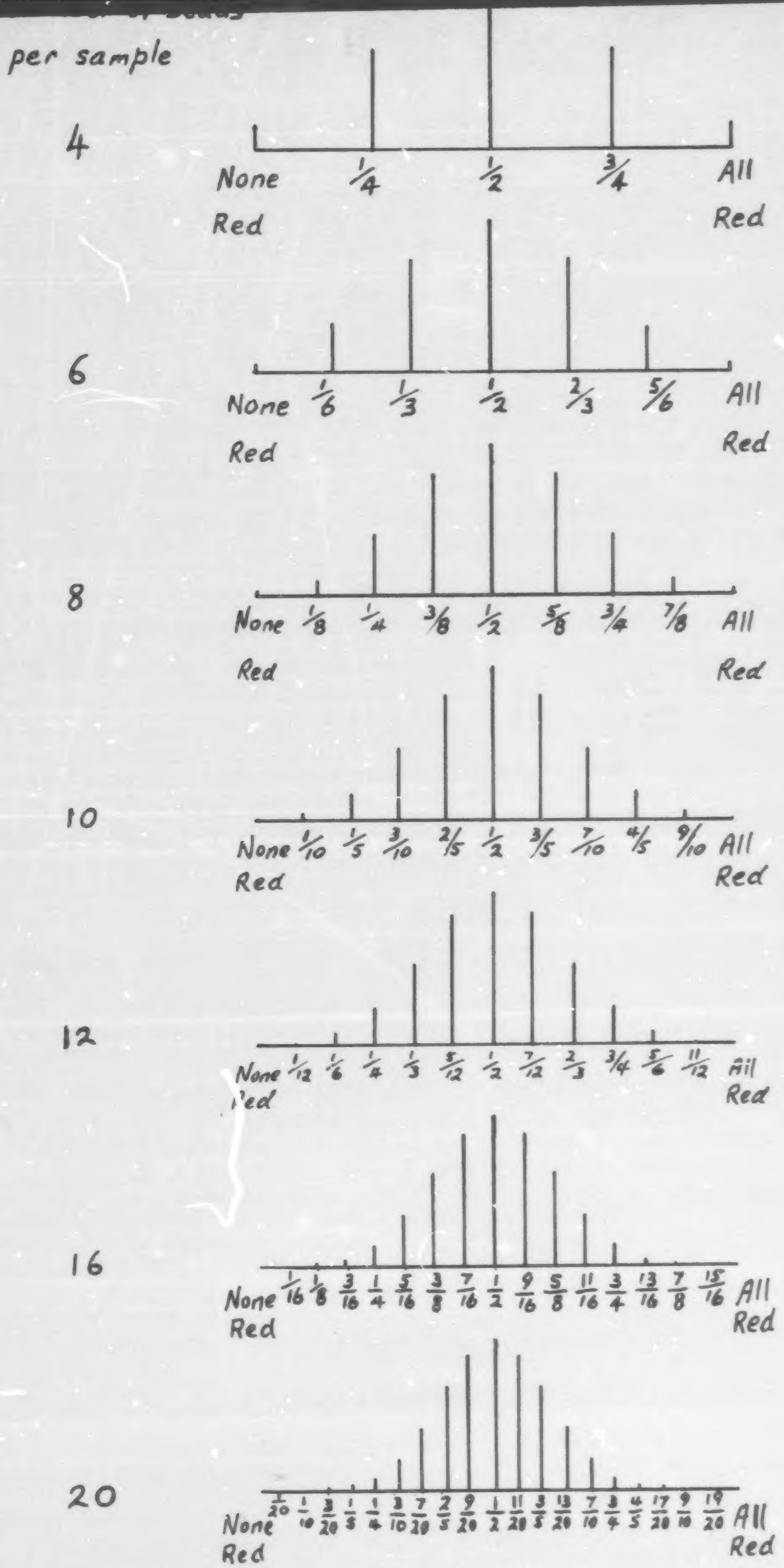
Some obvious but important observations can be made about these distributions. They are all unimodal (i.e. they have one hump) and they are perfectly symmetrical. The peak of the distribution, for every one of these sample sizes, coincides with the true proportion for the population. That is to say, the single most frequent type of sample is the one that gives a true picture of the population. (Half the beads in the sack are red, and the single most frequent type of sample, regardless of the sample size, is the type containing half red.) The next most frequent types are those that give a result close to 'half red' and the least frequent types are those that give the worst results ('none red' or 'all red').

As the sample size gets larger, the distribution takes on a distinct shape. If you joined together the tops of the vertical lines in the bottom diagram, and then smoothed out the corners, you'd get something like this. (I've lowered the baseline slightly in the right hand diagram, to show that, though the curve gets very close to the baseline, it never actually touches it.)





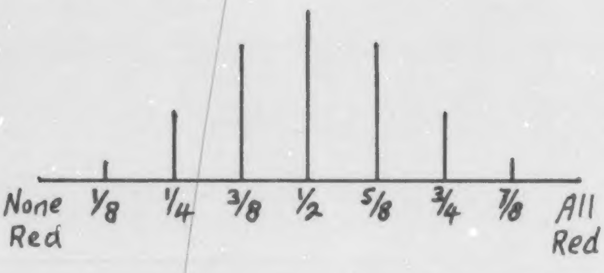
As the sample size gets larger, the theoretical distribution of results from samples of that size gets closer to the shape of this curve. The distribution that takes this shape is called 'the normal distribution' and the curve is called 'the normal curve'. The normal distribution crops up often in statistics. One reason for its importance is that it shows how a

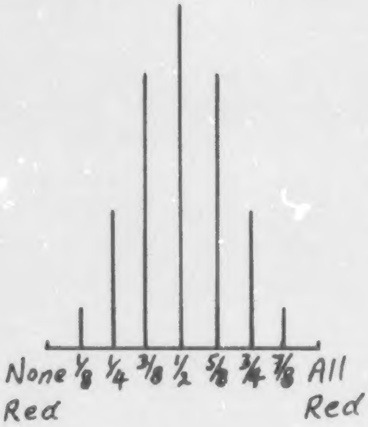

per sample
4
None Red
1/4
1/2
3/4
All Red
6
None Red
1/6
1/3
1/2
2/3
5/6
All Red
8
None Red
1/8
1/4
3/8
1/2
5/8
3/4
7/8
All Red
10
None Red
1/10
1/5
3/10
2/5
1/2
3/5
7/10
4/5
9/10
All Red
12
None Red
1/12
1/6
1/4
1/3
5/12
1/2
7/12
2/3
3/4
5/6
11/12
All Red
16
None Red
1/16
1/8
3/16
1/4
5/16
3/8
7/16
1/2
9/16
5/8
11/16
3/4
13/16
7/8
15/16
All Red
20
None Red
1/20
1/10
3/20
1/5
1/4
3/10
7/20
2/5
9/20
1/2
11/20
3/5
13/20
7/10
3/4
4/5
17/20
9/10
19/20
All Red

particular result ('proportion of red beads', in my example) would vary from one random sample to another, taken from the same population.

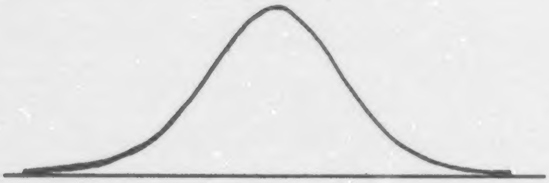
The normal distribution

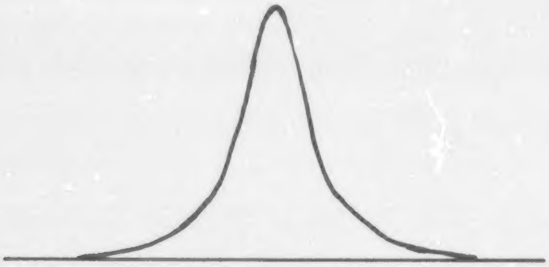
In order to explain an important feature of the normal distribution, I should first clear up a possible confusion about what is meant by the 'shape' of a distribution. Diagrams can be a bit misleading here, because the visual shape of a diagram depends partly on how you draw it. Take, for example, one of the earlier diagrams showing the theoretical distribution of a result from random samples of eight beads. On the left it is as I drew it before, with the base line measuring 8 cm and the centre line 2.5 cm; on the right I've made the base line 4 cm and the centre line 5 cm:



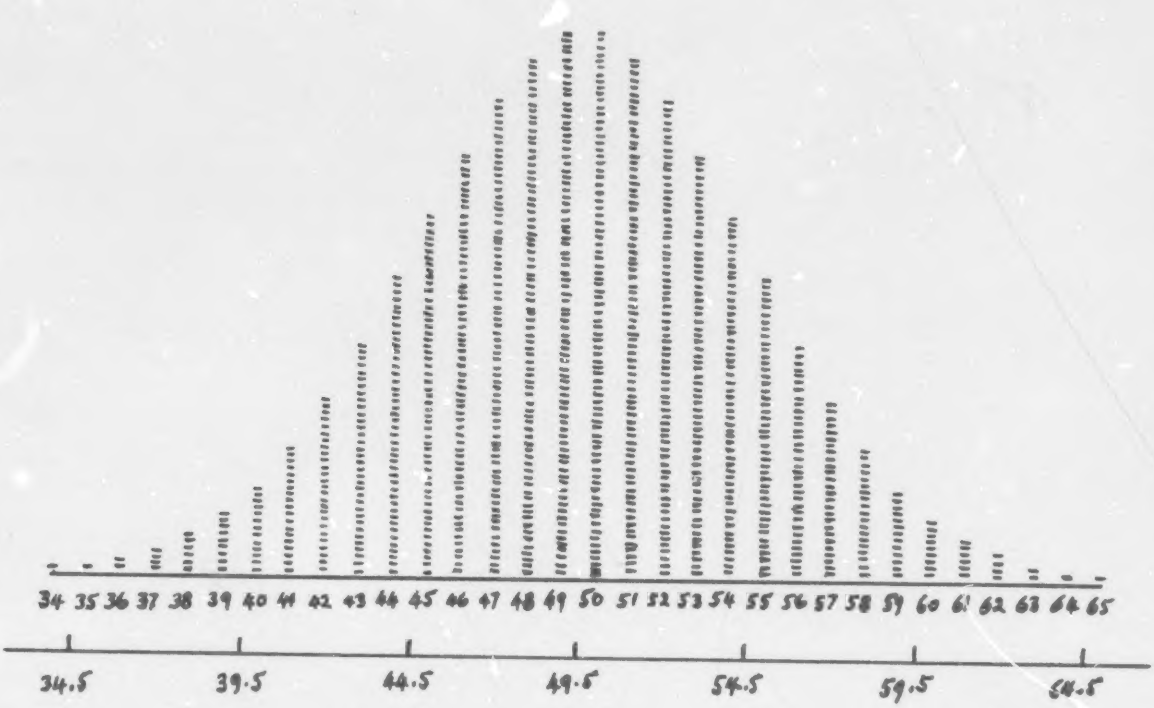


These are two different diagrams of the same distribution, though they might not look like it. What they have in common is the relative height of the lines; the relationship between the centre line and the next highest lines, for instance, is the same in both (the second highest are four-fifths the height of the centre line). The relationship between different parts of the distribution is the important thing. Thus it is that the following two diagrams both show the normal curve, even though they look different:





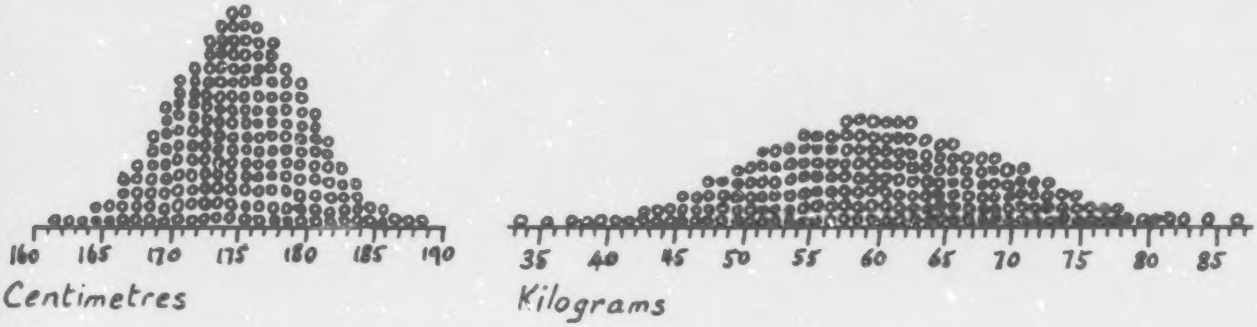
It is not immediately obvious how you would describe what it is that these two curves have in common. It is the standard deviation that provides the key. Returning to the sack of beads, imagine that you took a random sample of 99 beads and noted the number of them that were red, and that you repeated this procedure 1 000 times. You would end up with 1 000 'scores', each score being the number of red beads counted in one sample - a number between 0 and 99. Theory predicts that, if you carried on forever, these scores would form a normal distribution with 49 and 50 being the most common scores (50% is the true proportion of red beads in the sack and 50% of 99 is 49.5) and tailing off to each end. A real set of 1 000 samples would not produce exactly this pattern, but let's suppose that your 1 000 scores were distributed almost exactly as theory would predict, i.e. in a form closely resembling the normal distribution. Each dot in this diagram represents one score, i.e. the number of red beads in a single sample of 99:



Like any distribution of scores, this distribution has a mean, a median and a standard deviation. The mean is 49.5; the median is also 49.5; and the standard deviation is 5 (slightly over 5, actually, but 5 is near enough). Moving downwards from the mean, one standard deviation takes you to 44.5 (i.e. 49.5 - 5); moving upwards from the mean, it takes you to 54.5. If you count all the dots that come between these two points, i.e. all the scores from 45 to 54 inclusive, you'll find there are 682. In other words 68.2% of the scores, which is slightly over two-thirds, fall within one standard deviation (up or down) of the mean. If you move out to two standard deviations from the mean (all scores from 40 to 59 inclusive), you'll find that 954 of the scores (95.4%) fall within this range. And if you move out to three standard deviations from the mean, all but two of the scores (99.8%) fall within the range.

This is true of every set of scores that follows the normal distribution. A certain range on either side of the mean (expressed in terms of the standard deviation) takes in a certain proportion of the scores. This is true no matter what the scores represent and no matter what the values of the mean and the standard deviation are.

Suppose you took a random sample of 200 men aged 30, and measured the height and weight of each one. This would give you two sets of 'scores' - 200 measurements of height (in cm) and 200 measurements of weight (in kg). You would probably find that both sets of scores formed a normal distribution, approximately. Let's imagine that they resembled the normal distribution very closely and that you drew a diagram of them, like this:



In the left-hand diagram, each circle represents one man's height; in the right-hand diagram, each circle represents one man's weight. These 200 men do not differ very much in height, but they differ a lot in weight, so the right-hand diagram is flatter than the left-hand one. But both sets of measurements follow closely the normal distribution. The mean of the distribution of heights is 175 cm and the standard deviation is 5 cm, so 68% of the men have heights between 170 and 180 cm and 95% between 165 and 185 cm. In weight their mean is 60 kg and the standard deviation is 10 kg, so 68% have weights between 50 kg and 70 kg, and 95% between 40 kg and 80 kg.

Estimating probability

To recap, independent random samples drawn from the same population produce results that vary from one sample to another. If you took lots of samples, the sample results would themselves form a distribution. The types of sample that gave a good or fairly good picture of the population would occur more often than those that gave a bad picture. As the sample size gets larger, the distribution of results tends to adopt a distinctive pattern known as the normal distribution. If a set of scores has a normal distribution, then, if you know the mean and standard deviation, you know precisely how the scores are arranged around the mean.

Two more steps are needed to show how this underlies the calculation of confidence limits. I have shown that, for the sack of beads with half red and half blue, repeated sampling gives a normal distribution of results, as the sample size gets larger, and that the type of sample that occurs most often is the type that resembles the population, i.e. the type containing half red. It turns out that this is true regardless of the true proportion in the population. If, say, one quarter of the beads in the sack were red and you took repeated samples, the sample results would still form a normal distribution, this time with the most frequently occurring sample being the one containing one-quarter red.

Although the sack of beads is a very simplified example, the same principles apply to any random sampling. Suppose you took a random sample of 150 households to estimate the proportion of households in the entire country who owned a radio. If 1 000 survey organisations conducted this survey at the same time, their results would form a normal distribution. If, in fact, 40% of all households owned a radio, then most of these samples would give a result close to '40%', with just a few samples giving results rather far from '40%'.

How many would be close to '40%', and how many would be far from '40%', and how far would the furthest be? Because these results form a normal distribution, we could answer these questions precisely if only we knew the standard deviation. Well, we can calculate the standard deviation by the following formula:

$$s = \sqrt{\frac{pq}{n}}$$

$$s = \sqrt{\frac{40 \times 60}{150}} = \sqrt{\frac{2400}{150}} = \sqrt{16} = 4$$

So, about two-thirds (68%) of these sample results would fall between '36%' and '44%', 95% of them between '32%' and '48%', and almost none would be outside the range '28%' to '52%'.

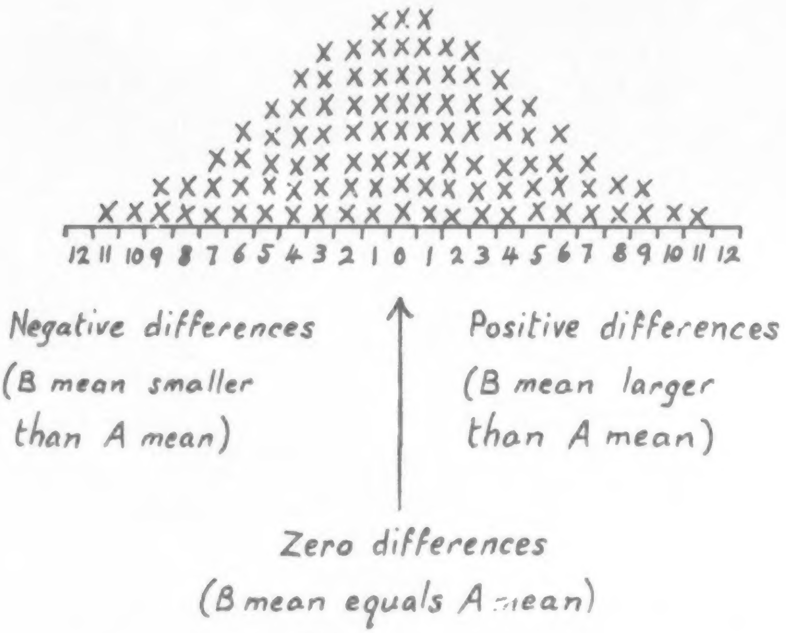
Does this formula look familiar? If you glance back to Appendix 1, you'll find that I gave something very similar as a formula for estimating the 'standard error'. The standard error is the standard deviation of the normal distribution of results that you'd get if you did your survey hundreds of times.

With both the sack of beads and the radio-owning survey, I've assumed that we knew the true proportion in the population and that we were taking hundreds of samples. In real life, you don't know the true proportion in the population - that's why you're doing a survey - and you only take one sample, not hundreds. You just turn the argument round.

Ninety-five per cent of all samples will give a result within two standard errors of the real figure. Putting it another way, there's a 95% chance that a sample result lies within two standard errors of the true figure. It follows that, for any particular sample result, there's a 95% chance that the true figure is not more than two standard errors away. If you could estimate the standard error from your sample result, this would tell you how close you probably were to the true figure. The formula I gave in Appendix 1 is the one you use to estimate the standard error from the results of a single random sample.

Some more steps are needed to show how these same principles form the basis of significance tests. Up to now I have talked about sample results that take the form of a proportion - the proportion of beads that were red, the proportion of households that owned a radio. A sample result can also take the form of a mean; for example, if you did a sample survey on cattle farming, you might calculate the mean cattle holding from your results, e.g. 'The mean holding was 14 cattle.' I won't go through the steps in the argument as I have for proportions, but it turns out that sample means also follow the normal distribution. If a thousand surveys each took a random sample of 100 cattle farmers, and the mean cattle holding was calculated from each survey, these 1 000 means would form a normal distribution, with the peak coinciding with the true mean for the population.

If you drew two independent random samples, each of 100 cattle-farmers, and calculated the mean cattle holding for each one, the figures would probably be slightly different; the mean for sample A might be 15 cattle and for sample B 17 cattle, a difference of two cattle (the B mean minus the A mean). Now, imagine that you repeated this many times, i.e. taking two independent samples and calculating the mean cattle holding from each one, and then calculating the difference between the two means. Sometimes the B mean would be higher (a positive difference), sometimes it would be lower (a negative difference) and sometimes the two means would be the same (zero difference). Generally the two figures would be close (giving a difference not far from zero), but occasionally the difference would be large. In short, these differences would form a distribution and, as you've probably guessed, this would be a normal distribution. The results of 100 pairs of surveys might look like this. Each cross here represents a difference between the results of two surveys:



Suppose you have given one version of a lesson to one group of students - group A - and another version to group B, then given them all the same test and then calculated that group B did better, on average, than group A. Is the difference statistically significant, i.e. what would be the probability of getting this difference in scores just by chance, when really one group was no better prepared for the test than the other? Imagine a large population of students and think of your two groups as independent random samples drawn from that population. If one version of the lesson was no more effective than the other, the two groups would be about equally well prepared for the test. But, just as two samples of cattle farmers would probably give slightly different results, these two groups of students would probably get slightly different marks on the test, just by chance.

Repeating the procedure many times - drawing two independent samples of students and calculating the difference between their mean scores - you'd get a normal distribution of differences. If you could calculate that the standard deviation of this distribution was, say, three marks, you'd know that there was a 68% chance of getting a difference of up to 3 between the mean scores and a 95% chance of getting a difference of up to 6 marks. Putting it the other way round, a difference of more than 6 marks would occur by chance only 5% of the time. If the difference between your two groups was, say, 7 marks, you could say that, if one lesson was no better than the other, you'd get such a result less than 5% of the time. You would conclude, therefore, that one of the lessons was probably more effective than the other.

The statistical problem, then, is to calculate the standard deviation of this normal distribution. Trying to explain how that is done, however, would take me into deep water. If you want to know more about it, consult one of the statistics books listed in Appendix 4. I hope, however, that I've explained the basic idea on which these tests are based. You regard your particular result (from a survey or from an experiment) as just one of a large number of possible results that you might have got. Given certain assumptions (about the randomness of the sampling and so on), you can deduce with mathematical precision how these results would be distributed. You can then say where your result is located in that distribution - whether it's one of those which would occur commonly or one that would occur only rarely.

Appendix 3 Costing

Imagine the policy-makers of a distance-teaching organisation trying to reach a decision. They might be faced with two alternative approaches to some educational problem and be wondering which to adopt. Or they might be deciding how to allocate resources among the different departments, or whether to repeat a project that they have done before. One question they are likely to ask is 'How much does it cost?' In terms of the examples I have just given, the questions would be, 'Is one of these approaches more expensive than the other?' 'How much is each department currently spending?' 'How much did this project cost when we did it before?'

You might expect that the accounts department, rather than the research department, should be called upon to consider the question, 'How much does it cost?' Of course the accounts people keep records of all the organisation's financial transactions, so that they can say where the money comes from and where it goes, and they produce a periodic summary - once a year at least - of the organisation's dealings over the previous period and of its current financial position. From this financial statement, you could see, for example, how much was being spent on rent, or fuel, or salaries or printing materials. But these records are not generally arranged so as to give a ready answer to a question such as, 'How much does it cost to produce a booklet?' Someone has to go through the accounts, picking out those portions of various expenditures which went on this particular piece of work. Often, it is the researcher who has to do this, though he will need the help of the accounts people, especially if he does not know much about accounting.

What to put in and what to leave out

Say you were asked to assess the cost of producing a correspondence course. You might list the items as follows:

Writer's salary, full-time, 24 months

Editor's salary, equivalent to 6 months

Artist/photographer, 3 months

Course typist, 12 months

Printer, 2 months

Printing materials (paper, ink, etc.)

Then you remember that the writer made a few trips around the country to discuss parts of the course with other teachers, and that the researcher helped to pre-test the course. And, of course, the writer, editor, artist and typist all used paper and other materials, so you add these to the list:

Writer's travel expenses

Researcher, 2 weeks

Writer's, editor's, artist's and typist's materials

So that's everything that went into the production of that course. Is it? In one sense, it is. It is a list of those items that were directly and obviously connected to producing the course. But there are many other items which are not on that list because they are not so obviously connected to that course. The writer had a desk in an office; the organisation had to pay rent for the office and had to pay for the heating, lighting and cleaning. The writer received a salary cheque every month; the organisation had to pay the accountant who made out these cheques. The machine that printed the courses used electricity, and the printer used soap to clean his hands. And so on. The organisation has to pay all sorts of bills, just in order to function. Accountants refer to these as 'overheads'. The correspondence course could not have been produced without these items, so some portion of the organisation's total overheads should be added to the cost of the course.

So far, I have been talking about those items on which the organisation has to spend its own money. Sometimes, there are costs which are borne by other organisations. For example, you might produce your own radio programmes and then take the tapes to the national radio service for broadcasting. If the national radio service does not charge you anything, then the broadcasting is 'free' to your organisation. But the broadcasting is not really free; someone has to pay for the upkeep of the transmitters and the electricity that they consume. It's just that, in this example, these costs are borne by the national radio service. Similarly the government's extension agents might distribute leaflets that you have produced; if they regard this as part of their regular duties, the distribution is 'free' to your organisation, but their time is still being paid for - by the government.

Should you include items like this in the costs of a piece of work? Do you include radio transmission costs, for example, even though your organisation does not have to pay them? It depends on why you are assessing the costs. In the examples I gave on the previous page, the policy-makers were wondering how to spend the money of their own organisation. Their question 'How much does it cost?' was really 'How much does it cost us?' For their purposes, you would not include costs which are borne by other organisations. If the policy-makers were wondering whether to give more or less money to their radio department, they would want to know what proportion of their own organisation's resources were being spent by the radio department. It would not matter to them if, say, transmission costs had doubled in the previous year since they don't have to pay them anyway.

In short, the organisation's policy-makers are, understandably, interested mainly in their own organisation's resources, so they want to know how much of those resources a piece of work costs. People in other positions, however, might take a different view. Suppose that the distance-teaching organisation is funded by an annual government grant, and that someone in the government's planning office is trying to decide what proportion of the education budget should go to the distance-teaching organisation. He might try to compare the costs of providing secondary education by distance teaching with the costs of providing it in schools. He would certainly include the costs of transmitting the distance-teaching radio programmes because the government has to pay those costs and he is interested in how much of the government's resources are going into the distance teaching.

Which items you include in the costs and which items you leave out depends on whose resources you are talking about. Since this book is intended mainly for people who are working inside distance-teaching organisations, I will confine myself to talking about the costs which are borne by the organisation itself. Items for which the organisation does not have to pay will not appear in the costs. I must immediately add a word of warning to this, which is that you cannot use these costs to compare distance teaching with other sorts of teaching, or to compare one distance-teaching organisation with another. Suppose that one organisation has to buy its own vehicles and then has to pay for fuel and servicing, whereas another organisation can borrow vehicles from the government's transport pool, paying nothing at all. Every time the first organisation costs a piece of work, it has to include something for transport, whereas the second does not. It would obviously be wrong to conclude that the second organisation was more economical.

A method of assessing costs

I ought to confess that the advice I am about to give, unlike the advice in the rest of the book, is second-hand; that is to say, I have never actually gone through this procedure myself. However, colleagues of mine have used this method and they tell me that it works. My reason for including it is that, if this advice had been given to me when I was at LDTC, I would have made some attempt at analysing costs. As it was, I never did because I couldn't see how to go about it.

The method assumes that your organisation keeps fairly detailed records of its financial transactions, so that you could find out, for example, how much the courses editor has been paid in the last year, how much the new printing machine cost and how much has been spent on ink and paper. Not all organisations keep such accounts. If the organisation was part of a government ministry, for example, it might order its equipment and supplies on government requisition forms rather than pay for them directly out of its own bank account. In such a case, you'd have to make enquiries to find out how much had been spent, using whatever records were available (there might be duplicate copies of the forms, for instance) and making guesses for things you couldn't find out about.

It will become clear as I describe the method that, even if you have detailed accounts, you have to make guesses at various points in the calculations. This means that the result is not very accurate. It is important to remember this because, when dealing with numbers, it is possible to end up with a result that appears to be very precise, e.g. 'The cost per booklet is 43.72 cents.'

But this appearance of precision is misleading. In getting to that figure, you will have made various guesses, any of which might be seriously inaccurate. As in presenting the results of a social survey, you should try to indicate how accurate you think your estimates are; you might say 'The cost per booklet is between 35c and 50c.'

The method begins by drawing up a table which presents the organisation's total expenditure for a particular period, such as the previous financial year. This table might be useful in itself, but its main value is that it provides you with some key figures which you can then use in costing particular pieces of work. Drawing up the table might require a few days' work, but you need to do it only once a year. Once you have these key figures, it is fairly easy to cost particular pieces of work.

I will give examples in terms of Lesotho's unit of currency, the Maloti, abbreviated to M; M1 = 100 cents. At 1981 rates, the Maloti had a value somewhere between the American dollar and the pound sterling.

One year's expenditure

Before I describe the table into which you put the year's total expenditure, I should explain more clearly what I mean by 'one year's total expenditure'. What you need is a figure for the total financial resources used up during a year. These financial resources will be expressed as amounts of money. But they are not exactly the same as actual amounts of money paid out during the year. You cannot simply add up all the cheques paid during the year; it is a bit more complicated than that.

One thing you must do is to convert capital expenditure into an amount per year. 'Capital' means items such as furniture, vehicles and machines which are expected to provide useful service for quite a long time - more than a year, anyway. Say you bought a large printing machine in the last financial year, at a cost of M15 000. If you included the entire M15 000 in the printing costs for that single year, then printing during that year would appear very expensive and the printing in the following years much cheaper. That would obviously be misleading. What you do is to assess how long the machine will give useful service - the manufacturer might tell you, or professional printers. If you expected your machine to last for five years, you could simply divide the purchase price by five and say that the machine was costing M3 000 per year for five years.

Another way to distribute the costs of capital over several years is to say that it loses a certain proportion of its value each year. People who have bought or sold a second-hand car will be familiar with this idea already. A car which was bought, new, for M5 000 might lose 30% of its value each year. So, after one year, it would be worth 70% of its original value, i.e. 70% of M5 000, which is M3 500. After another year it is worth 70% of M3 500, i.e. M2 450. Putting it another way, you could say that the cost of this item of capital was M1 500 for the first year, M1 050 for the second year, and so on. Here again, you might have to consult specialists in order to find out the appropriate percentage to deduct each year. (Accountants refer to this percentage

- the figure of 30% in the above example of the car - as the 'depreciation rate'.) You need to do these calculations for capital only if you own the capital; if you hired your printing equipment, instead of owning your own, you would just calculate the hiring charges for the year.

Make sure that, so far as possible, you include all expenditure for the year in question, and also that you do not include expenditure for the years before or after. Whether or not you actually paid the bills during the year is irrelevant. Take fuel bills, for example. If you received your last quarterly bill two months before the end of the financial year, you would make an estimate of the amount that you would have to pay for those last two months, even though you might not have paid it yet.

Similar problems can arise if you have bought large amounts of stock. For example, if you bought an exceptionally large amount of paper for printing, but you used only a small proportion of it during the year in question, then you should put down the cost of the paper that was actually used, not the cost of the whole lot. Conversely, you might have used some paper that you had bought the year before. Again, you put down the cost of the paper that was used during the year in question, even though it had already been paid for.

It might seem reasonable that you should deduct your income and the value of your assets from the expenditure. For example, say you have printed 500 copies of a correspondence course during the year. At the end of the year you have sold 50 at M20 each and you have the other 450 in stock. You might argue that, if you take account of the money that you will get back from these courses (M1 000 for the 50 already sold and M9 000 for the remaining 450 which you will sell eventually), the courses will have cost very little. But this is jumping the gun. At the moment, we are still trying to find out how much it actually did cost to produce these 500 copies. Whether or not this cost will later be offset by income from selling the courses is a separate question. So, at this stage of the procedure, do not deduct income and the value of assets from the expenditure.

All the things I have mentioned so far will appear in the 'Income and Expenditure Accounts' prepared by the accountant. However, you will have to consult the detailed accounts for the year rather than just taking figures from the annual summary. The reason is that accountants group things together in their summaries, partly for convenience and partly to present the affairs of the organisation in the best light. The money which the organisation contributes to the staff's Christmas party, for example, will probably not be labelled as such, but will be grouped with some other things under a heading such as 'Public relations' or 'Personnel recreation and training'. You will need to go into more detail to get a good estimate of how much was spent during the year, and what it was spent on.

Finally I must qualify something I said earlier. I said that I would confine the discussion to the costs which the organisation bears, ignoring items which are provided 'free' to the organisation. An exception to this is those items which the organisation does not pay for now but which it will have to pay for in the near future. For example, say the radio producer is a volunteer from overseas whose salary is paid directly by the volunteer agency and does not pass through the organisation's accounts at all. His services are 'free' to the organisation. But, if he is to be replaced soon by a local person whose salary will not come directly from overseas, it might give a more realistic picture of the organisation's costs if you included a salary for the producer even though, at present, the organisation is not actually paying it.

The table of the year's total expenditure

If, with the help of the accounts department, you can arrive at a figure for the total financial resources used up during a year, you can then put this expenditure into a table. The first part of the table looks like this (the numbers in the boxes are just to help me to explain the procedure):

	A Salaries	B Materials	C Other
Course writing and editing	5	15	25
Student services	6	16	26
Rural education	7	17	27
Literacy project	8	18	28
Research	9	19	29
Printing	10	20	30
Radio	11	21	31
Transport	12	22	32
	↓	↓	↓
Sub totals	13	23	33
Other	14	24	34
Totals	2	3	4
Grand total		1	

The total financial resources used up during the year go into box number 1. You divide this amount between salaries (box 2), materials (box 3) and the remainder (box 4).

Then you go to the top of the table and fill in column A (salaries). You take each department in turn and you put down the salaries of those people who give all their time, or almost all of it, to that department. Into box 5 go the salaries of the course writers and the course editor, for example. (I have included eight departments in the table, which would be appropriate for costing the work of the LDTC; people in other organisations might put in different departments.)

There may be some staff members who divide their time between several departments. The artist , for example, might work for the course-writing section, the rural education section and the literacy project. In that case you would try to get from him an estimate of the proportion of his time that he spends in each department and you would divide his salary accordingly.

When you have filled in boxes 5 to 12, add them up to get the salaries subtotal (box 13). Almost certainly you will find that this is less than the amount in box 2. There are members of staff, such as the director, the accountant, and the cleaner, who do not work in any of the departments in the table (or, to put it another way, they work for all these departments). So you subtract box 13 from box 2 to get box 14 (other salaries).

Now you turn to column B (materials) and do the same sort of thing. Be careful that you don't include the same costs twice. For example, if the printing section printed 300 questionnaires for the research section, put this down either in box 19 or in box 20, but not in both. Again, you will find that box 23 (materials subtotal) contains a smaller amount than box 3. Subtracting one from the other gives box 24. This is for materials such as floor polish and instant coffee which are used by all departments.

Then onto column C (Other). Put down the cost of items other than salaries and materials which can be ascribed directly to each department. The postage costs of corresponding with the students, for example, would go into box 26; the depreciation cost of the printing equipment would go into box 30; the depreciation cost of the vehicles, plus their fuel and maintenance costs, would go into box 32. Again, subtract box 33 from box 4 to get box 34. This could include the rent for the building, electricity bills, the contribution to the staff Christmas party and all sorts of other items.

You have now filled in the first part of the table. Next you add another column to the table, so that it looks like this:

	A Salaries	B Materials	C Other	D General overheads
Course writing and editing	5	15	25	36
Student services	6	16	26	37
Rural education	7	17	27	38
Literacy project	8	18	28	39
Research	9	19	29	40
Printing	10	20	30	41
Radio	11	21	31	42
Transport	12	22	32	43
	↓	↓	↓	↑
Sub totals	13	23	33	
Other	14	24	34 →	35
Totals	2	3	4	
Grand total		1		

You add together boxes 14, 24 and 34 to give box 35. This is the total expenditure which you were not able to assign to any particular department. Call it 'general overheads'.

The next step is to assign these general overheads among the departments. It is probably not worthwhile to attempt immensely detailed calculations in order to decide what proportion of the general overheads should go to each department. Simply assign them on some rough-and-ready basis. For example, you could divide them according to the number of staff members in each department.

If these eight departments had 24 staff members altogether, and if the course-writing and editing section contained six of these, then you would assign a quarter of the general overheads (6 as a proportion of 24) to box 36. In this way you would split box 35 between boxes 36 to 43.

The final steps require one more column to the table, like this:

	A Salaries	B Materials	C Other	D General overheads	E Total overheads
Course writing and editing	5	15	25	36 →	44
Student services	6	16	26	37 →	45
Rural education	7	17	27	38 →	46
Literacy project	8	18	28	39 →	47
Research	9	19	29	40 →	48
Printing	10	20	30	41 →	49
Radio	11	21	31	42 →	50
Transport	12	22	32	43 →	51
	↓	↓	↓	↑	↓
Sub totals	13	23	33	↑	52
Other	14	24	34 →	35	
Totals	2	3	4		
Grand total	↖	↑ 1	↗		

Simply add together columns C and D for each department. So, boxes 25 and 36 are added together to give box 44 and so on. The boxes in column E give you the amount of expenditure, per department, that was composed of everything except the department's salaries and materials. You could call these the 'total overheads'. If you want, you can add up boxes 44 to 51 to give box 52. As a check on your arithmetic, box 13 + box 23 + box 52 should equal box 1.

As I said earlier, the table might be of some value in itself. It shows the relative size of salaries, materials and overheads for each of these departments. It also shows the allocation of the organisation's resources among these departments. However, for the purpose of costing particular pieces of work, its main value is that the figures in boxes 44 to 51 provide a basis for allocating overheads. One final step is required for this.

You need to express the total amount of work performed by each department over the year in question, in units of some kind. The most convenient units for most of the departments will be 'person-days', a 'person-day' being a day of work done by one person. Say the research department consists of a research officer and two assistants and that for the year in question they have all worked five days a week for 48 weeks. The total work of the research department, expressed in person-days, is 5 x 48 x 3, which comes to 720 person-days.

You can now express the overheads of the research department in terms of person-days. Suppose that the total overheads of the research department (box 48 in the table) came to M3 000. Then you could express this as M4.17 per person-day (i.e. M3 000 divided by 720). In itself this is not a particularly interesting or meaningful figure, but it will be useful later on.

Other departments might use units other than person-days. The printing department might keep records of the time that the printing machines are running. Suppose that the department has two machines and that they have been running for 6 hours a day, 5 days a week for 45 weeks in the year (i.e. actually running, not being cleaned or repaired). You could express the work of the printing department in terms of 'machine-hours': 6 x 5 x 45 x 2 = 2 700. If the department's overheads (box 49) were M9 500, you could express this as M3.52 per machine-hour.

Similarly, you might express the radio department's work in terms of 'studio hours', i.e. the hours during which the studio was actually in use. And you might express the transport department's work in terms of 'vehicle-kilometres', i.e. the number of kilometres travelled by the department's vehicles over the year.

The end product of all these calculations is eight figures. These are the total overheads per person-day (or per machine-hour or per vehicle-kilometre) for each department. They are not very meaningful in themselves, but they are needed in the next section.

Costing one piece of work

Now I can return to the problem of costing a particular piece of work, such as the production of a correspondence course. I will explain the procedure in terms of another table which partly resembles the one I used in the last section. To make it clear that it is a different table, I will refer to the boxes by means of letters instead of numbers.

Course writing and editing	a	i	--- person days at ---	q
Student services	b	j	--- " " " ---	r
Rural education	c	k	--- " " " ---	s
Literacy project	d	l	--- " " " ---	t
Research	e	m	--- " " " ---	u
Printing	f	n	--- machine-hrs at --	v
Radio	g	o	--- studio-hrs at ---	w
Transport	h	p	--- kilometres at ---	x
Subtotals	tot sal	tot mat		tot over
Grand total		Grand total		

You begin by finding out how much time each member of these departments has spent on this piece of work. If they have not kept records of their time, they will have to make estimates. It is better, however, if you can get them to keep time sheets. A time sheet is a simple form on which, at the end of each week, each staff member records how much of his time he has spent on the various pieces of work he has been involved in. One week's time sheet for the artist, for example, might look like this:

Staff member ..Artist......... Week beginning 4/Mar/79
Maths course 3 days
Nutrition booklet 1½ days
Family planning poster ½ day

If staff members fill in these time sheets regularly, you can calculate fairly accurately how much of each person's time has gone into a particular piece of work.

You then calculate each person's time in terms of his salary, like this:

Course writer	2 years at M3 500 per year	= 7 000
Editor	6 months at M350 per month	= 2 100
Artist	3 months at M200 per month	= 600

When you have added up the department's total salary bill for this piece of work, you enter it in the table. The salary bill for the input of the course-writing and editing department, for example, goes into box a. If one department had not been involved at all in this piece of work, you would enter 0 in the relevant box.

Having filled in boxes a to h, you move on to materials. Here again, if people have not kept records, you will have to ask them to make estimates of the amount of materials that have gone into this piece of work. But it is better if they keep stock cards on which they record the materials that they use. One of the printer's stock cards might look like this:

Type of material 2-SHEET BOARD A4 SIZE, WHITE

Date	In	Out	Stock in hand	Job
5/4/81	30 reams		30 reams	
8/4/81		2 reams	28 "	Maths course
15/4/81		17 reams	11 "	Nutrition booklet

By consulting the stock cards, you find out what materials were used on the piece of work in question. You find out from the accounts department how much the various items cost. Then you calculate the materials bill, for this piece of work, for each of these departments, and enter these amounts in the table (in boxes i to p).

Then the overheads. For the first five departments (i.e. from course writing and editing down to research), you would find out the number of person-days that have gone into this piece of work. (In fact, you have already collected the necessary information for the purpose of calculating the salary bill.) Take the research department as an example. Say you find that the research officer has spent ten days on this piece of work, and one of the research assistants has spent fifteen days; the total person-days are 25. From the calculations described in the last section, you know the figure for total overheads per person-day for the research department. Say this figure was M4.17. For this piece of work, you calculate the research department's overheads as 25 person-days at M4.17 per person-day, which comes to M104.25. You enter this figure in box u.

Similarly, you find out from the printing department the number of machine-hours that have gone into this work; from the radio department you find the number of studio-hours; and from the transport department you find the number of vehicle-kilometres. You might have to rely on estimates of these, but it is better if people keep log books. The printer would record the number of machine-hours devoted to various pieces of work; the radio officer would record the number of studio-hours and the vehicle drivers would record the number of kilometres.

All these figures, along with the figures calculated in the last section, enable you to fill in boxes q to x. Finally, you just add up the columns and then add the three subtotals together to give the grand total. This is the total financial resources, overheads included, which the organisation expended on that piece of work. It is the answer to the question we began with - 'How much does it cost us?'

Unit costs and extra costs

Suppose you calculated that a small, pilot literacy project had cost M2 100. Is that a large or a small amount? It depends partly on how many students there were. If there had been 50 students you could express the cost as M42 per student. Costs expressed as so-much-per-something are called 'unit costs'.

Often, there is more than one way of calculating unit costs. To continue the example of the literacy pilot project, you could take the number of students who enrolled for the course, or the number that progressed to a certain point, such as completing half the course, or the number who completed the full course and passed a test at the end. If only five students completed the course, someone could argue that the cost-per-student was M420, not M42. Organisations will be tempted to select the unit cost that presents their projects in the best light, but it is more honest to present a range of unit costs - so much per student who enrolled, so much per student who completed the course, and so on.

When people are considering the costs of a piece of work, they should have the full picture, not just the grand total or one particular unit cost. The reason is that their decisions will affect different parts of the cost in different ways. Suppose that they are looking at the costs of producing radio programmes, and suppose that these costs seem rather high. If they were looking only at the total cost, they might think they could economise simply by producing fewer programmes, but this might not be true. Perhaps the main component of these costs is the high depreciation cost of some expensive studio equipment. They would not reduce this cost by using the equipment less.

Similarly, some parts of the costs might be more worrying, when looking to the future, than other parts of the costs. An organisation which relies on donor agencies for part of its income is, to some extent, at the mercy of the donor agencies' policies. Despite the unemployment problems that beset developing countries, donor agencies are often happier to give equipment than to pay wages; they will donate a collating machine rather than pay the wages of two people to do collating, even though the costs, over five years or so, might be the same. An organisation in that position would be more worried about the salaries component in the cost than about the depreciation of the equipment.

Say you produced 500 copies of a leaflet at a total cost of M280, giving a unit cost of 56c per leaflet. This includes the cost of the writing, illustrating and pre-testing, as well as the printing. If you were then asked to print 500 more, would this second lot also cost M280? Obviously not. You don't need to do the writing, illustrating and pre-testing again. Just printing another 500 copies might cost as little as M40, so the unit cost of the second lot would be 8c. This second unit cost is the extra cost of producing one more leaflet, granted that all the work up to that point has already been done and paid for. It is sometimes called the 'marginal cost'.

A strong economic argument for distance teaching is that, though you spend a lot of money producing the materials, you can still keep down the cost per student by using those materials to teach many students, and you can teach many students because the marginal cost per student is quite low. It is expensive, for example,to produce a new correspondence course, but the marginal cost of teaching each extra student - postage and tutor fees and so on - is low, so you can afford to enrol many students for the course. Using the same course to teach many students justifies spending a lot of money on producing the course in the first place. If you use radio, the marginal cost per student is actually zero. That is to say, all your expenditure goes into producing and broadcasting the programme; it is the same whether you have 500 listeners or 5 000.

Marginal costs are often the ones that policy-makers are most interested in. If you think that the results of a pilot literacy project are encouraging and you are wondering whether to repeat it on a larger scale, the unit cost of the pilot project (M42 per enrolled student, or whatever), though of some interest, is not really the figure you want. Perhaps a high proportion of the pilot expenditure went on developing the materials and testing them on only a few students. The expanded project will use these same materials with many more students, so the cost-per-student in the expanded project should be much lower than in the pilot.

A final word of advice is that you should always check unit costs back against the total. It is easy to make mistakes in these calculations that can lead to your result being as much as ten times too high or too low. Suppose that, according to your calculations, your organisation apparently spent M450 last year on each of its correspondence students. If you had 500 students, the total cost must have been well over M200 000. Checking this figure against last year's total expenditure, you could see at a glance if it was feasible. If, say, your total expenditure on all activities was only M150 000, you'd know your figure was completely wrong.

Appendix 4 Further reading

I feel I ought not to recommend a book unless I've read it myself, or at least a large portion of it. This principle cuts down drastically the number that can compete for a place in this list. There may well be many other books that deserve to be recommended as highly as these, or even more highly, but, if there are, I don't know them.

References on social research

MOSER, C.A. and KALTON, G. *Survey methods in social investigation,* second edition. Heinemann Educational Books, London, 1971. (550 pages)

HOINVILLE, Gerald and JOWELL, Roger and Associates, *Survey research practice.* Heinemann Educational Books, London, 1978. (228 pages)

WEBB, Eugene J. and others, *Unobtrusive measures: nonreactive research in the social sciences.* Rand McNally and Company, Chicago, 1966. (226 pages)

HALL, Budd L. 'Participatory research: an approach for change', *Convergence* (an international journal of adult education) Vol. VIII, No. 2, 1975, pages 24-32.

O'BARR, William M, SPAIN, David H. and TESSLER, Mark A. *Survey research in Africa: its applications and limits,* Northwestern University Press, Evanston, 1973. (350 pages)

The first is the standard British textbook on social surveys. Not easy reading, but thorough. The second is more about the practical, organisational problems of conducting surveys. Both are about surveys in western, urban societies, but much of what they say applies equally to work in developing countries. The third is a collection of methods that have been used to avoid the problem of reactivity, i.e. the problem of people behaving in a special way when they know they are being studied.

Many people hold the view that social research methods of the kind I have described in this book, especially social surveys, are inappropriate for use in developing countries. The article by Hall presents some of these arguments. The book by O'Barr and others is a collection of papers by various authors; some of them criticise the use of survey methods in Africa, and others give examples from several African countries of how they have adapted survey techniques to cope with research problems. The book also contains a large bibliography on research in Africa.

References on experimental design and action research

CAMPBELL, Donald T. and STANLEY, Julian C. *Experimental and quasi-experimental designs for research*. Rand McNally College Publishing Company, Chicago, 1966 (84 pages)

CAMPBELL, Donald T. 'Reforms as experiments', pages 233-261 of *Readings in evaluation research* edited by CARO, Francis G. Russell Sage Foundation, New York, 1971.

CAMPBELL, Donald T. 'Administrative experimentation, institutional records and nonreactive measures', pages 257-291 of *Improving experimental design and statistical analysis*, edited by STANLEY, Julian C. Rand McNally Company, Chicago, 1967.

HORNIK, Robert C. 'Useful evaluation designs for evaluating the impact of distance learning systems: methodology', *Educational broadcasting international*, March 1976, pages 6-10.

The first of these papers covers the subject very thoroughly. Unfortunately, it is written in an abstract style which makes it hard to read. This sentence is fairly typical: 'Regression effects are usually a negatively accelerated function of elapsed time and are therefore implausible as explanations of an effect at O_5 greater than the effects at O_2, O_3 and O_4.' Even reading it in context, I'm not sure what this means. However, the parts I can understand are very good. The other three papers are easier, but less exhaustive.

References on statistics

BLALOCK, Hubert M. Jnr. *Social statistics,* revised second edition. McGraw-Hill Kogakusha Ltd. 1979 (584 pages)

LANGLEY, Russell, *Practical statistics for non-mathematical people,* David and Charles (Publishers) Ltd, 1968 (400 pages)

HUFF, Darrell, *How to lie with statistics,* first published 1954. Victor Gollancz, London. Also available from Penguin Books (124 pages)

MEEK, Ronald L. *Figuring out society,* Fontana paperback, 1972 (236 pages)

The first is a standard American textbook, intended primarily for university students. It is good but not easy; readers need a good grasp of basic mathematics. The second has less theory and gives more 'how-to-do-it' instructions. The other two are descriptions of how quantitative methods (i.e. methods that involve measurement and calculation) and statistics are used, and often misused, to describe and explain social patterns and events. Both books are entertaining as well as instructive.

References on other topics

MAGER, Robert F. Preparing instructional objectives. Fearon Publishers Inc., Palo Alto, California, 1962 (60 pages)

A short and clear self-instructional text about defining objectives so as to make educational efforts more amenable to evaluation.

POPHAM, W. James. Criterion-referenced measurement. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1978 (260 pages)

This book describes the difference between tests that find out how much someone has learnt about something ('criterion-referenced tests') and those that find out whether someone has more or less ability than another ('norm-referenced tests'). The first kind are the ones you need for evaluating distance teaching. Popham then shows how you can construct such tests. The context he has in mind is the American school system, but the principles apply elsewhere.

Agricultural Extension and Rural Development Centre, University of Reading, Berkshire, U.K. Action research and the production of communication media, 1974 (32 pages)

This is the report of a workshop held in India on using research in the production of instructional materials on agriculture. They use the expression 'action research' for what I prefer to call 'pre-testing'. It is clear and well illustrated.

KLARE, George R. The measurement of readability, Iowa State University Press 1963 (Text to page 190, bibliography pages 191-328). The Flesch Reading Ease score that I described in chapter 12 is only one of many measurements of readability. This book describes many more and also explains how they were devised.

JAMISON, Dean T., KLEES, Steven J. and WELLS, Stuart J. The costs of educational media: guidelines for planning and evaluation, Sage Publications, 1978 (255 pages)

A useful book if you want to know more about costing.

GOWERS, Sir Ernest, The complete plain words, second edition, Penguin Books, 1973 (332 pages)

A book about how to write good English. It was written originally for civil servants but it applies just as well to people writing research reports. I don't agree with him on every detail - we differ on the use of the word 'data' for instance - but his general advice is excellent.

MORRIS, Lynn Lyons and FITZ-GIBBON, Carol Taylor. *Program evaluation kit.* Sage Publications, Beverly Hills and London, 1978.

This kit consists of eight books, varying in length from about 80 to 180 pages. It gives clear and practical 'how-to-do-it' advice to people evaluating educational efforts and it is written more for beginners than for experts. It covers many of the topics that I have treated in this book, generally at greater length. The main difference is that the kit is intended primarily for people doing evaluation within the American school system. It would be useful for people who wanted more advice on topics that I have treated rather briefly, such as defining educational objectives or measuring attitudes. You can buy the books separately but they are cheaper if you buy the full kit. The eight titles are:

1. Evaluator's handbook
2. How to deal with goals and objectives
3. How to design a program evaluation
4. How to measure program implementation
5. How to measure attitudes
6. How to measure achievement
7. How to calculate statistics
8. How to present an evaluation report

Appendix 5 References

Introduction	p. 2	A more detailed history of LDTC is given in: Murphy, P. The Lesotho Distance Teaching Centre: Five Years' Learning; International Extension College Broadsheet 16, 1981.
Chapter 1	p. 8	The survey of 250 rural adults was reported in: Understanding print: a survey in rural Lesotho of people's ability to understand text and illustrations, LDTC, 1976. Copies are available from Lesotho Distance Teaching Centre, price M3. (See below for details of how to obtain copies of LDTC reports.)
	p. 8	'The postal services in Lesotho', a one-page report by LDTC, 1974.
	p. 8	The survey of organisations was reported as: 'Notes on non-formal education in Lesotho: first draft', LDTC, 1974.
	p. 9	'Housewives' choices', LDTC, June 1976.
	p. 9	'Survey of private candidates', LDTC, January 1977.
	p. 10	The test of line drawings and photographs was part of the survey reported in Understanding print - see above.
	p. 10	'A test on the best way to present a correspondence lesson', LDTC, October 1976.
	p. 10	'Learning from radio', LDTC, 1976.
	p. 11	Fuglesang, Andreas Applied communication in developing countries, Dag Hammarskjold Foundation, 1973.
	p. 11	'Poultry keeping in rural Lesotho', LDTC, October 1975.
	p. 12	Attitudes to family planning in Lesotho, LDTC, 1977. Copies available from LDTC, price M3.
	p. 13	'Family planning pamphlet: pre-test/evaluation', LDTC, September 1976. The final version of the family planning pamphlet was published jointly by the LDTC and the LFPA with the title 'Thero ea Malapa K'eng?'

Chapter 1	p. 13	The final version of the crochet booklet was published by LDTC as Buka ea ho Korocha.
	p. 13	Examples of monitoring reports are: 'Progress of LDTC students', published annually by LDTC since 1976.
	p. 15	'The survey on broadcasting', LDTC, 1976. 'Listenership among secondary school students, LDTC, 1976. 'Listenership among LDTC's students to LDTC's broadcasts for JC students', LDTC, 1976.
	p. 15	'CRS booklet evaluation', LDTC, January 1977.
Chapter 2	p. 20	See the references above (p. 15) and also: 'Evaluation of radio support for private candidates', LDTC, 1978.
	p. 21	The observation of shopping and the cataloguing of reading matter were reported first in: 'A brief survey of the uses of literacy and numeracy in Lesotho', LDTC, May 1976, and later as part of Reading, writing and arithmetic in Lesotho, LDTC, 1979 (available from LDTC, price M3)
	p. 21	'Distance teaching and road safety', LDTC, June 1977.
Chapter 3	p. 29	Lefatshe La Rona - Our Land: the report on the Botswana government's public consultation on its policy proposals on tribal grazing land. Ministry of Local Government and Lands, Republic of Botswana, September 1977. See also a series of papers entitled 'Technical notes on the tribal grazing land policy consultation campaign', Evaluation Unit, Botswana Extension College.
	p. 31	'Study of pre-natal and child care practices in rural Lesotho', LDTC, 1976.
Chapter 4	p. 36	'Clothes-making in rural Lesotho', LDTC, January 1976.
	p. 38	'Testing a family planning pamphlet', LDTC, May 1975.
	p. 41	Cambridge elementary statistical tables, Cambridge University Press. There are also tables in: Blalock, Hubert M. Social Statistics, McGraw Hill Kogakusha, 1979. Many other textbooks of statistics contain statistical tables.

Chapter 5	p. 58	'Evaluation of radio support for private candidates', LDTC, 1978.
Chapter 6	p. 69	Understanding print, LDTC, 1976.
Chapter 7	p. 94	Nie, Norman H. and others, SPSS: Statistical package for the social sciences, second edition, McGraw-Hill Book Company, 1975 (675 pages). An easier introduction to SPSS is given in: Klecka, William R. and others, SPSS primer, McGraw-Hill Book Company, 1975 (134 pages).
Chapter 8	p. 101 p. 106	Understanding print, LDTC, 1976.
	p. 101 p. 107	Reading, writing and arithmetic in Lesotho, LDTC, 1979.
Chapter 10	p. 133	The research at the Hawthorne plant was reported in: Roethlisberger, F.J. and William J. Dickson, Management and the worker, Harvard University Press, 1939.
	p. 136	'Housewives' choices', LDTC, June 1976.
	p. 137	Understanding print, LDTC, 1976.
	p. 139	Learning from a booklet: an experiment with individuals and groups in Lesotho, LDTC, 1978. Copies available from LDTC, price M3.
Chapter 12	p. 168	Farr, J.N., Jenkins, J.J. and Paterson, D.G. 'Simplification of Flesch reading ease formula' Journal of applied psychology, Vol. 35, No. 5, October 1951, pages 333-337.
	p. 170	The extracts from correspondence courses were taken from: Bookkeeping and commerce for Junior Certificate, LDTC and Production techniques 1A, Royal Melbourne Institute of Technology, Dept. of External Studies.
	p. 171	Spaulding, S. 'A Spanish readability formula', Modern Language Journal, 40, December 1956, pages 433-441.
Chapter 13	p. 174	'Testing a family planning pamphlet', LDTC, May 1975.
	p. 182	Taylor, Wilson L. 'Cloze procedure: a new tool for measuring readability' Journalism quarterly 30 (Fall 1953) pages 415-433.

Chapter 13	p. 183	The short passage used as an example is taken from: *The soil - how the soil is made up*, Better Farming Series No. 7. Food and Agriculture Organisation, Rome 1970.
Chapter 14	p. 194	'Evaluation of radio support for private candidates', LDTC, 1978.
	p. 195	'Progress of LDTC students', LDTC, annually from 1976.
	p. 198	Two memoranda from the National Extension College, Cambridge, England: 'Examination results and correspondence courses', September 1966 'Examination results and home study courses', October 1967.
Chapter 15	p. 204	'Evaluation of TBRDP/LDTC training and support programme for VDPAs', LDTC, December 1976.
	p. 206	'CRS booklet evaluation', LDTC, January 1977.
Chapter 16	p. 222	*Learning games*, LDTC, June 1977. Copies available from LDTC, price M3.
	p. 222	*Lefatshe La Rona.* For full reference, see above.
	p. 224	*Learning from a booklet*. See above.

Copies of LDTC's research reports are available, while stocks last, from:

Lesotho Distance Teaching Centre
PO Box 781
Maseru 100
Lesotho, Africa

A complete list of the Centre's research reports is given in LDTC's progress report, published twice a year.

People who want copies of reports should write to LDTC listing the ones they want and enclosing a bank draft or international money order for M3 (three Maloti - the currency of Lesotho)* to cover surface-mail postage. Most of the reports themselves are free, except for some longer ones for which a further charge is made of M3 each.

---

* Or R3 (three Rands - the currency of the Republic of South Africa, which is also accepted in Lesotho).

Appendix 6 Addresses of organisations

Here are the names and addresses of some organisations which may be able to help with suggestions or information about distance teaching and other people's experience of practical research.

The *International Extension College*, whose address is at the front of this handbook, publishes practical handbooks and a series of Broadsheets on Distance Learning. It also offers an information service and runs a training course for people working in distance-teaching institutions.

The *Clearinghouse for Development Communication*, 1414 22nd Street, NW, Washington, DC 20037, USA, publishes the *Development communication report*, a free newsletter which appears four times a year. It is a centre for materials and information on applications of communication technology to development problems.

The *Institute for International Studies in Education* of Michigan State University (513 Erickson Hall, Michigan State University, East Lansing, Michigan 48824, USA) includes a Non-formal Education Information Centre which publishes a newsletter, *The NFE Exchange*, twice a year.

For information on African society and development generally, a good source is the African Institute for Economic and Social Development (INADES), BP 8008, Abidjan, Ivory Coast. Its documentation department offers a question and answer service, with the aim of linking ordinary people in Africa with modern sources of information. It has a very large library, mainly consisting of material on Africa, agriculture, politics and economics.

There is a *Steering Committee for an International Institute for Distance Learning*, based at the Open University, Walton Hall, Milton Keynes, MK7 6AA, England. Its interest is mainly in university-level learning, and it has a growing resource centre.

The *Institute for Communication Research*, Cypress Hall, Stanford University, Stanford, California 94305, USA, offers a Master of Arts degree in communication and development, and also considers students for admission to MA programmes in communication research (specialising in health, nutrition or other applications of communication theory and research), journalism, film and broadcasting.

The University of Reading Agricultural Extension and Rural Development Centre, London Road, Reading, RG1 5AQ, England, publishes a Bulletin three times a year, with particularly good book reviews and descriptions of rural education projects.

The East-West Centre, 1777 East-West Road, Honolulu, Hawaii 96848, USA, is an educational institution which exists to promote better relations and understanding between the United States and the nations of the Pacific through co-operative study, training and research. It responds to requests for information, and offers grants to people seeking solutions to problems of mutual consequence to East and West.

Unesco (the United Nations Educational, Scientific and Cultural Organisation) publishes an enormous range of books, periodicals and papers on education. You can obtain a complete catalogue of its publications from the Office of the Unesco Press, 7 place de Fontenoy, 75700 Paris, France.

World Education, 1414 Sixth Avenue, New York, NY10019, USA, works to strengthen the capabilities of agencies working directly with educationally disadvantaged people in Asia, Africa, Latin America and the United States. Its Reports Magazine is published three times a year and distributed free to readers professionally engaged in nonformal education for adults. Others pay an annual subscription of US $5.

Finally, two organisations in Latin America. The Centro Latinoamericano de educación de adultos is at Av. Providencia 2093, Cásilla 16.417 - Correo 9, Santiago, Chile, and the Instituto Latinoamericano de la Comunicacion Educativa is at J. Luis Vives 200, Apartado Postal 94-328, México 10, D.F., Mexico. Both publish regular information bulletins.

Appendix 7 Statistical tables

Table 1 Random numbers

The numbers are grouped in fives to make them easier to read, but you can treat them as continuous strings of numbers. If, for example, you were taking numbers in groups of three, starting on the fifth line down, you could take 241-926-861 and so on.

63255	01320	42392	66212	20100	80563
54333	22877	90674	25248	98392	77359
12955	71715	40052	18732	76051	60417
79458	38428	76451	17686	89340	82415
24192	68613	94593	76488	73076	76339
37232	68664	14294	46827	82713	82629
40434	81216	97362	54652	21151	10696
15439	12012	96143	03091	15326	32587
61455	29377	61201	49512	39872	95165
48750	73111	18238	53157	85087	07399
25697	37188	48531	75514	88522	66810
52549	19444	07891	87802	57597	84694
93078	58272	88520	38477	64295	05772
04922	73499	25765	72233	65363	73269
53639	28911	98613	35419	13752	51605
92546	04003	29588	68700	29907	53438
54901	61154	45196	92910	28503	11773
48677	41171	01481	77718	73805	81783
52902	60156	59051	13770	28415	54399
33087	87424	46382	91165	89639	19657
03619	06050	89212	72589	87565	50641
42422	40627	26785	75823	91112	63931
64068	43328	70275	54967	73712	39123
34575	88158	08641	93504	87621	43144

Table 2 Values of χ^2 (chi-squared)

Degrees of freedom	Significance levels			
	20% (one chance in five)	10% (one in ten)	5% (one in 20)	1% (one in 100)
1	1.64	2.71	3.84	6.63
2	3.22	4.61	5.99	9.21
3	4.64	6.25	7.81	11.34
4	5.99	7.78	9.49	13.28
5	7.29	9.24	11.07	15.09
6	8.56	10.64	12.59	16.81
7	9.80	12.02	14.07	18.48
8	11.03	13.36	15.51	20.09
9	12.24	14.68	16.92	21.67
10	13.44	15.99	18.31	23.21
11	14.63	17.28	19.68	24.73
12	15.31	18.55	21.03	26.22
13	16.99	19.81	22.36	27.69
14	18.15	21.06	23.68	29.14
15	19.31	22.31	25.00	30.58
16	20.47	23.54	26.30	32.00
17	21.62	24.77	27.59	33.41
18	22.77	25.99	28.87	34.81
19	23.90	27.20	30.14	36.19
20	25.04	28.41	31.41	37.57

You calculate the numbers of degrees of freedom from your table of results by the formula: (rows - 1) × (columns - 1). Then you look along the corresponding row of this table to see where your value of χ^2 comes. For instance, from a table of results with four degrees of freedom, a value of 8.02 for χ^2 would be significant at the 10% level (being larger than 7.78) but not at the 5% level (being smaller than 9.49).

Table 3 Values of t

Degrees of freedom	Significance levels 20% (one chance in five)	10% (one in ten)	5% (one in 20)	1% (one in 100)
1	3.078	6.314	12.706	63.657
2	1.886	2.920	4.303	9.925
3	1.638	2.353	3.182	5.841
4	1.533	2.132	2.776	4.604
5	1.476	2.015	2.571	4.032
6	1.440	1.943	2.447	3.707
7	1.415	1.895	2.365	3.499
8	1.397	1.860	2.306	3.355
9	1.383	1.833	2.262	3.250
10	1.372	1.812	2.228	3.169
11	1.363	1.796	2.201	3.106
12	1.356	1.782	2.179	3.055
13	1.350	1.771	2.160	3.012
14	1.345	1.761	2.145	2.977
15	1.341	1.753	2.131	2.947
16	1.337	1.746	2.120	2.921
17	1.333	1.740	2.110	2.898
18	1.330	1.734	2.101	2.878
19	1.328	1.729	2.093	2.861
20	1.325	1.725	2.086	2.845
21	1.323	1.721	2.080	2.831
22	1.321	1.717	2.074	2.819
23	1.319	1.714	2.069	2.807
24	1.318	1.711	2.064	2.797
25	1.316	1.708	2.060	2.787
26	1.315	1.706	2.056	2.779
27	1.314	1.703	2.052	2.771
28	1.313	1.701	2.048	2.763
29	1.311	1.699	2.045	2.756
30	1.310	1.697	2.042	2.750
40	1.303	1.684	2.021	2.704
60	1.296	1.671	2.000	2.660
120	1.289	1.658	1.980	2.617
over 120	1.282	1.645	1.960	2.576

If you have calculated t from two sets of independent scores, the degrees of freedom are the total number of scores minus two, e.g. for groups of 15 and 17 there are (15 + 17 - 2) = 30 degrees of freedom.

If you have dependent sets of scores, the degrees of freedom are the total number of pairs minus one.
You find where your value of t comes in the above table. For instance, with 25 degrees of freedom, a value of 1.6 for t would be significant at the 20% level (being larger than 1.316), but not at the 10% level (being smaller than 1.708).

10% level of significance

Number of students in the larger group

Number of students in the other group

	9	10	11	12	13	14	15	16	17	18	19	20
1											0	0
2	1	1	1	2	2	2	3	3	3	4	4	4
3	3	4	5	5	6	7	7	8	9	9	10	11
4	6	7	8	9	10	11	12	14	15	16	17	18
5	9	11	12	13	15	16	18	19	20	22	23	25
6	12	14	16	17	19	21	23	25	26	28	30	32
7	15	17	19	21	24	26	28	30	33	35	37	39
8	18	20	23	26	28	31	33	36	39	41	44	47
9	21	24	27	30	33	36	39	42	45	48	51	54
10	24	27	31	34	37	41	44	48	51	55	58	62
11	27	31	34	38	42	46	50	54	57	61	65	69
12	30	34	38	42	47	51	55	60	64	68	72	77
13	33	37	42	47	51	56	61	65	70	75	80	84
14	36	41	46	51	56	61	66	71	77	82	87	92
15	39	44	50	55	61	66	72	77	83	88	94	100
16	42	48	54	60	65	71	77	83	89	95	101	107
17	45	51	57	64	70	77	83	89	96	102	109	115
18	48	55	61	68	75	82	88	95	102	109	116	123
19	51	58	65	72	80	87	94	101	109	116	123	130
20	54	62	69	77	84	92	100	107	115	123	130	138

5% level of significance

Number of students in the larger group

Number of students in the other group

	9	10	11	12	13	14	15	16	17	18	19	20
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

The smaller the value of U, the more significant the result. If you had, for example, 17 students in one group and 11 in the other, a value of 54 for U would be significant at the 10% level (being smaller than 57) but not at the 5% level (being larger than 51).

Table 5 Values of T for the Wilcoxon matched-pairs signed-ranks test

Number of pairs	Level of significance		
	10%	5%	1%
6	2	0	-
7	3	2	-
8	5	4	0
9	8	6	2
10	10	8	3
11	13	11	5
12	17	14	7
13	21	17	10
14	25	21	13
15	30	25	16
16	35	30	20
17	41	35	23
18	46	40	28
19	53	46	32
20	60	52	38
21	67	59	43
22	74	66	49
23	82	73	55
24	92	81	61
25	101	89	68

For a given number of pairs, a value of T is significant if it is equal to or smaller than the one given in this table. For instance, with 20 pairs, a value of 47 for T would be significant at the 5% level (being smaller than 52), but not at the 1% level (being larger than 38).

Tables 2-5 of this appendix are reproduced or adapted, with permission, from Hubert M. Blalock Jnr., Social statistics, 2nd edition, McGraw-Hill Kogakusha Ltd., Tokyo, 1972.

Table 6 The Flesch reading ease score

Average sentence length (words per sentence)	Number of monosyllables per 100 words																									
	84	82	80	78	76	74	72	70	68	66	64	62	60	58	56	54	52	50	48	46	44	42	40	38	36	34
9	94	90	87	84	81	78	74	72	68	65	61	58	56	52	49	45	42	40	36	33	29	27	23	20	17	13
10	93	89	86	83	80	77	73	71	67	64	60	57	55	51	48	44	41	39	35	32	28	26	22	19	16	12
11	92	88	85	82	79	76	72	70	66	63	59	56	54	50	47	43	40	38	34	31	27	25	21	18	15	11
12	91	87	84	81	78	75	71	69	65	62	58	55	53	49	46	42	39	37	33	30	26	24	20	17	14	10
13	90	86	83	80	77	74	70	68	64	61	57	54	52	48	45	41	38	35	32	29	25	23	19	16	13	9
14	89	85	82	79	76	72	69	67	63	60	56	53	50	47	44	40	37	34	31	28	24	22	18	15	12	8
15	88	84	81	78	75	71	68	66	62	59	55	52	49	46	43	39	36	33	30	27	23	21	17	14	11	7
16	87	83	80	77	74	70	67	65	61	58	54	51	48	45	42	38	35	32	29	26	22	20	16	13	10	6
17	86	82	79	76	73	69	66	64	60	57	53	50	47	44	41	37	34	31	28	25	21	19	15	12	9	5
18	85	81	78	75	72	68	65	63	59	56	52	49	46	43	40	36	33	30	27	24	20	18	14	11	8	4
19	83	80	77	74	71	67	64	61	58	55	51	48	45	42	39	35	32	29	26	23	19	17	13	10	7	3
20	82	79	76	73	70	66	63	60	57	54	50	47	44	41	38	34	31	28	25	22	18	16	12	9	6	2
21	81	78	75	72	69	65	62	59	56	53	49	46	43	40	37	33	30	27	24	21	17	15	11	8	5	1
22	80	77	74	71	68	64	61	58	55	52	48	45	42	39	36	32	29	26	23	20	16	14	10	7	4	
23	79	76	73	70	67	63	60	57	54	51	47	44	41	38	35	31	28	25	22	19	15	13	9	6	2	
24	78	75	72	69	66	62	59	56	53	50	46	43	40	37	34	30	27	24	21	18	14	12	8	5	1	
25	77	74	71	68	65	61	58	55	52	49	45	42	39	36	33	29	26	23	20	17	13	11	7	4		
26	76	73	70	67	64	60	57	54	51	48	44	41	38	35	32	28	25	22	19	16	12	10	6	3		
27	75	72	69	66	63	59	56	53	50	47	43	40	37	34	31	27	24	21	18	15	11	9	5	2		
28	74	71	68	65	62	58	55	52	49	46	42	39	36	33	30	26	23	20	17	13	10	8	4	1		
29	73	70	67	64	61	57	54	51	48	45	41	38	35	32	29	25	22	19	16	12	9	7	3			
30	72	69	66	63	60	56	53	50	47	44	40	37	34	31	27	24	21	18	15	11	8	6	2			
31	71	68	65	62	59	55	52	49	46	43	39	36	33	30	26	23	20	17	14	10	7	5	1			
32	70	67	64	61	58	54	51	48	45	42	38	35	32	29	25	22	19	16	13	9	6	4				
33	69	66	63	60	57	53	50	47	44	41	37	34	31	28	24	21	18	15	12	8	5	2				
34	68	65	61	59	56	52	49	46	43	40	36	33	30	27	23	20	17	14	11	7	4	1				
35	67	64	60	58	55	51	48	45	42	38	35	32	29	26	22	19	16	13	10	6	3					
36	66	63	59	57	54	50	47	44	41	37	34	31	28	25	21	18	15	12	9	5	2					
37	65	62	58	56	53	49	46	43	40	36	33	30	27	24	20	17	14	11	8	4	1					
38	64	61	57	55	52	48	45	42	39	35	32	29	26	23	19	16	13	10	7	3						

You have calculated the number of monosyllables per 100 words and the average sentence length. You find the column for the number of monosyllables and the row for the sentence length. For instance if you had 78 monosyllables per 100 words and an average sentence length of 20 words, the RE score would be 73. If your number of monosyllables is an odd number, you take a score between the two that are shown. For example, with 83 monosyllables per 100 words and a sentence length of 9, you take a score between 90 and 94, i.e. 92.

This table is reproduced, with permission, from 'Simplification of Flesch reading ease formula' by J.N. Farr, J.J. Jenkins and D.G. Paterson, Journal of applied psychology, Vol. 35, number 5, October 1951, pp 333-337.

Index

accounts, financial, 283, 285-8
accuracy of results,
    acceptable level of, 101, 105, 111-12, 124, 145-7, 234-5, 249
    spurious appearance of, 143, 285-6, 296
action research, 21, 24, 137, 202
age groups, 54-5, 85-6, 101
analysis
    of pre-test results, 180-2
    of research results, 33-4, 109-24, 236-7
anchor items, 213
artists
    producing instructional materials, 5, 10
    involvement in pre-testing, 12, 16, 68, 172, 174-5, 180
attitudes, 11, 38, 185-6
attribute, 112
audience
    characteristics, 27-8
    preferences, 9
    size, 206
    see also students
average, 152, 153, 242

back-translation, 66
bar-chart, 117, 196
base total, 122, 142
baseline survey, 207, 239
before-and-after design
    in evaluation, 207-10, 213, 215
    in experiments, 128
behavioural objectives, 163-6, 176, 204, 299
bias
    in evaluation, 228-9, 230, 235
    in instructional materials, 221
    in interviewing, 33, 68, 103-6
    in marking tests, 130
    in samples, 99-101, 109-12
bimodal, 157, 160
bookkeeping
    correspondence course in, 15
    tests of, 216
booklets
    as action research, 137-8
    evaluation of, 15, 39, 51-2, 204-7, 210, 212, 214-16, 232
    experiments with, 136-7, 139-40
    pre-testing of, 13, 24, 173
    sales of, 24, 187
    topics for, 9
brackets in formulas, 246
branching in questionnaires, 59-61, 63, 71-2, 76

calculator, for statistics, 241
campaigns, evaluation of, 14, 35, 203, 207-11, 213, 218-22, 239
capital, financial, 286
card-puncher, 90-3
card-reader, 90
case, in survey analysis, 151, 242
cassette recorders, 11, 32, 163
category variable, 151-3, 241
cells in table, 254
census data, 44-6, 101
chance, *see* probability
checking survey results, 98-108
chi-squared test, 242, 252-9, 308
childcare, research into, 24, 31
chivvy-letters, experiment with, 126-7
closed questions, 52-3
cloze testing, 182-5, 304
cluster samples, 43-4, 44-6, 249-50
coding, 56, 77-82, 152-3, 181
coding frame, 79
comments
    from colleagues, 236-7
    from experts, 166-7, 221, 228
    from students, 185-6, 198, 221, 225
comparisons in evaluation, 200-1, 206-7, 223-4, 227
comprehension tests, 175-80
computers, 56, 89-96, 97, 106, 258
confidence limits, 41, 145-7, 249
confidential information, 49-50, 65, 71
consistency check, 106-8
contamination, 211
continuity correction, 259
control groups
    in evaluation, 210-12
    in experiments, 128, 129-31
cookery booklet, 15, 39, 206-7, 214
co-operative societies, 29, 39-41
correspondence courses
    costs, 283-4, 296
    evaluation, 15, 203, 204, 213, 220
    monitoring, 13-14, 187-201, 222
    postal system, 8
    see also students
correspondence lessons
    design, 10, 126-8, 131-2, 158-61
    expert comment, 166-7
    feedback, 196-8
    objectives, 163-6
    pre-testing, 12, 177-80, 186, 235
    readability, 167-71
costs
    of distance teaching, 220, 222-4, 283-96, 299
    of research, 16-17, 26, 227, 239
counter-sorter, 95-6, 97
counting variable, 112
criteria
    in evaluation, 198-201, 202-3, 218, 219
    in pre-testing, 165-6, 180
crochet booklet, 13, 139-40, 173, 224

crosstab/cross-tabulation, 77, 82-3, 87-9, 107, 112-14, 118, 142

'data', use of the word, 74 (footnote)
data-processing, 74-97
    methods compared, 96-7
    straight from questionnaires, 75-7
    with a computer, 89-95
    with a counter-sorter, 95-6
    with edge-clipped cards, 82-6
    with squared paper, 86-9
    with transfer cards, 77-82
degrees of freedom, 257
dependent scores, 260-1
dependent variable, 112
depreciation rate, 287
design of instructional materials, see materials design
design of research, 20-6, 234-5
    see also evaluation, experiments, questionnaires, surveys, tests
developing countries, 1, 4, 225
diagrams
    in instructional materials, 132
    in presenting research results, 117-21, 196, 208-9
directors of distance-teaching organisations, 5, 19, 229, 234
disc, for data storage, 94
discussion as a research method, 21, 30-4, 225
distance teaching, 1
    see also booklets, campaigns, correspondence courses, leaflets,
        radio, visual aids
distractors in multiple-choice questions, 178
distribution, statistical, 156-7, 159-60, 273-82
documentation centres, 11, 25, 305-6
documents, use of, 24-5
donor agencies, 227, 229, 231, 234, 237, 295
'don't know' replies in social surveys
    analysis, 101-3
    coding, 54, 77-9
drawings, see pictures
dropout rate, 191-6, 213-14
dummy questionnaire for training interviewers, 71

edge-clipped cards, 82-6, 97, 188
editors of instructional materials, 5, 10, 126-8, 131-2, 163-71
English
    readability of, 167-71, 182-5
    translating from, 12, 63-6
enrolment form, 189
ethical issues in research, 4, 30, 71, 134, 211, 214, 217-18
evaluation, 202-19, 220-32
    books about, 298-300
    by outsiders, 229-32
    designs, 206-12, 225-7
    need for tact, 233-8
    questions used in, 206, 213, 215-17
    usefulness of, 14-15
    using exam results, 200-1
evening classes, 133, 135-6, 200
exam results, 198-201

expected results, in chi-squared test, 252-4
expenditure, see costs
experimental group, 128
experiments, 21, 125-40
    books on design, 298
    comparable groups, 129-31
    external validity, 131-6
    internal validity, 126-9
    real-life, 125-6, 136-40
    statistical analysis, 157-60
experts
    on instructional materials, 166-7
    on research, 16, 25, 47, 123
exploratory research, 23, 34
external validity of experiments, 131-6

failure forms, 70
family planning
    attitudes to, 12, 28, 33, 38
    education in, 9, 28-9, 186
    knowledge of, 114-17
feedback, 187-8, 196, 198
fertiliser, sales of, 209-10
fieldwork in a survey, 67-73
film shows, 210-11, 213
filters in questionnaires, see branching
first aid booklet, 215-16
flipcharts, see visual aids
forced-choice questions, 59
formative evaluation
    explanation of the term, 14, 227
    see also evaluation, pre-testing of instructional materials
formulas, how to read them, 242-7
frequency count/distribution, 76, 109, 117

generalising from results, see external validity of experiments
government
    approval of surveys, 72
    statistics, 24, 99-101
group discussion, see discussion as a research method
group learning, 15, 30, 139-40, 224
    see also radio-learning groups

Hawthorne effect, 133, 135, 303
health education, 203, 207, 211
histogram, 118
hypothesis, 125, 128

illustrations, see pictures
illustrators, see artists producing instructional materials
inadequate information
    as a coding category, 75-6, 93
    in analysing and presenting results, 101-3, 144
    in chi-squared tests, 258
independent scores, 260-1

independent variable, 112
indirect measurement in evaluation, 216-18
input to a computer, 90, 95
internal validity in experiments, 126-8, 134-6, 208
interviewers,
  approach to respondents, 69
  bias, 33, 68, 103-6
  expenses, 72-3
  instructions, 69-71
  insurance, 72
  number of, 68
  pay, 72-3
  personal qualities, 31-2, 67-8
  role in pilot survey, 71
  sampling by, 44, 46-7, 69, 101
  supervision of, 72-3
  training, 69-71
  translating questionnaires, 64
interviewing
  in group discussions, 32-3
  in social surveys, 50, 60, 62, 70
item analysis, 181

judgement in evaluation, 218-19, 220-32, 236-7
  evaluator's own, 228

KAP surveys, 11-12
key-punch, 90

language
  of instructional materials, 12, 167-71, 182-5
  of questionnaires, 64-6, 72
leaflets
  audience for, 8
  costs, 295-6
  evaluation, 205, 210
  pre-testing, 12, 36, 164, 173, 176-7, 186
Lesotho, 2
  rural life in, 27
  sample of households in, 44
Lesotho Distance Teaching Centre
  description of, 2-3, 301
  research reports, 301-4
Likert scale, 165
literacy
  campaigns, 222, 225, 295
  research into, 8, 21, 23-4

Maloti, unit of currency, 286, 304
Mann-Whitney U test, 264-7, 310
marginal cost, 296
matching
  in evaluation, 212
  in experiments, 130, 260-1

materials design, 10-11
    assessment of, 163-71, 221
    see also booklets, correspondence lessons, leaflets, pictures, radio, visual aids
mean, 153-4, 242, 278, 281
mean deviation, 154-5
mean time between worksheets, 193-4
measurement variable, 151-3, 241
measuring string, 106
media, audience for, 8
    see also booklets, correspondence courses, leaflets, radio, visual aids
median, 153-4, 278
missing data, 101-3
mode, 157
monitoring, 13-14, 187-201, 222
multi-coding, 82, 93-4, 143
multi-stage sampling, 43-4, 44-6
multiple-choice questions, 130-1, 177-80, 182

negatives in questions, 177, 179
newsletter for agricultural project, 204
non-contact, 69, 99
non-parametric, 264
non-random samples, 39-41, 47, 147
non-response, 70, 99
non-sampling error, 145-7
normal distribution, 275, 277-9, 281-2
number games, experiment with, 135-6
numeracy, research into, 21, 29
nutrition, 142, 174, 206-7, 216-17

objections to research in developing countries, 4, 297
objectives
    of instructional material, 163-6, 299
    of projects, 203-4, 219, 226
observation as a research method, 21, 27-30, 217, 225
observed results, in chi-squared test, 252-3
open questions, 52-3, 82, 176, 180
organisations providing advice on distance teaching, 11, 25, 305-6
overheads, 284, 290

pamphlets, see leaflets
parametric, 264
participant observation, 29-30
participatory research, 297
pass rate in examinations, 199
pencil-and-paper tests, 174, 177-80
percentages, 99-101, 113, 142-5
    calculation of, 142
person-day, unit used in costing, 292
photographs, 10-12
phrasal verbs, 170
pictures, 10-13, 53, 137-8, 174-5, 180
pie-chart, 117

pilot
    projects, 139-40
    surveys, 71-2
policy guidance, 7-10, 18-20, 233-6, 239-40
population, sample from a, 99
postal surveys, 63
postal system, test of, 8
postcards, 8, 136
post-coding, 82, 93
posters, *see* visual aids
post-tests, *see* pre-tests
poultry keeping, survey of, 11, 39, 50-3, 74, 78-81, 144-5
practical research, *see* research
preamble to a questionnaire, 49-50
precoding, 56-7, 81, 82, 104
pre-testing of instructional materials, 172-86
    analysing results of, 180-2
    as a small survey, 36, 46
    importance of, 12-13, 172
    involvement of materials designers in, 68
    methods of, 173-4, 176-80, 182-6
    not pre-test of students, 128-9, 175-6
    preceded by assessment of materials, 163
    questions used in, 53, 174, 176-80
    see also booklets, correspondence lessons, leaflets, pictures, radio
pre-tests
    in evaluation, 213
    in experiments, 128, 135-6, 162, 261
probability, 40, 146-7, 148-50, 158, 161, 252, 254-5, 272-82
program, for computer, 94
protractor, 117
punched card, 90

quantitative, 35, 48
questionnaires, 48-66
    basic principle of, 35
    counting results from, 75-7
    design of, 48-66, 80, 104, 185, 206
    dummy, 71
    for computer analysis, 92-4
    self-completed, 49, 63-4, 185
    to correspondence students, 198
questions in questionnaires
    closed and open, 52-3, 63-4, 82
    forced-choice, 59
    leading, 57-8
    notes for, 50-2
    on delicate topics, 58
    pre-categorised, 53-6, 104
    pre-coded, 56-7, 81
    unanswerable, 57
    vague, 57
    see also evaluation, pre-testing of instructional materials

radio
    audience, 8, 13, 19, 22, 187, 240, 249
    costs, 223, 284, 295-6

  - evaluation, 20, 58, 193-4, 204-7, 210-12, 223, 233, 238-9
  - pre-testing, 12, 163-4, 172-4, 186
  - style of programmes, 10-11, 12
  - support for correspondence courses, 14, 15, 20, 193-4
- radio-learning groups, 29, 212, 218-19, 220, 222
- radio producers, 5, 12, 16, 172, 238-9
- random
  - assignment of experimental subjects, 129-30
  - assignment in evaluation, 210
  - numbers, 41, 307
  - position in multiple-choice questions, 179
  - sampling, 39-42, 272-82
- range check, 106
- reactivity, 30, 133, 217, 297
- readability
  - of instructional materials, 167-71, 173, 182-5, 299
  - of research reports, 123, 237-8
- Reading Ease (RE) score, 168-71, 303, 312
- record cards for correspondence students, 190, 192
- recording
  - discussions, 32
  - observations, 29
- recording sheet, 29
- records, use of, 21, 187-201, 209-10, 217-18
- refusals in surveys, 70, 72
- relations with other organisations, 8, 237
- reports of research, 121-3, 229, 236-8
- representative sample, 34-5, 39-41, 47, 99
- research
  - commissioned for the wrong reasons, 18-20
  - costs, 16-17, 26, 227, 239
  - methods (summary), 20-2
  - planning, 18-26, 234-5
  - practical as opposed to academic, 1, 15-17, 24-6, 134-5, 161, 227-8
  - priority over action, 139-40, 238-9
  - scale, 15-17, 25-6, 35-6, 123-4, 227
  - value, 1, 7-17, 227-8, 239-40
- researcher
  - as interviewer, 31, 68
  - contacts with other countries, 25, 237, 305-6
  - freedom of speech, 229
  - knowledge of languages, 66
  - personal opinions, 221, 226, 228-9
  - qualifications, 16, 238
  - relations with colleagues, 15, 18-20, 22-4, 165, 229, 230, 233-9
  - role in distance-teaching organisation, 1, 15-16, 22-4, 165, 172, 187-8, 228, 233-4
  - role in fieldwork, 67, 73
- respondents, 49
  - educational level, 63
  - giving false answers, 107
  - information on, 61-2
  - names of, 49, 130-1
- response rate, 98-9
- reweighting, 109-12
- rival explanation/hypothesis, 117, 127-8, 208, 215

- sample size, 46-7, 146-7, 150, 249, 255

sampling, 35, 38-47, 99
    bias, 99-101, 109-12
    by interviewers, 44, 46-7, 69, 101
    example of, 44-6
    theory, 272-82
sampling error, 145
sampling fraction, 42
schoolchildren
    as experimental subjects, 126-36, 185
    compared with correspondence students, 200
secondhand information, 38-9, 59
self-check exercises, experiments with, 126-8, 131-2
self-completed questionnaires, 49, 63-4, 185
semi-structured interview, 32
serial numbers
    on questionnaires, 49, 72-3, 78, 81, 91-2
    on record cards, 126
service-agency, 8, 9
shape of a distribution, 156-7, 159-60
short-answer questions, 177
side-effects, evaluation of, 225
significance level/test, see statistical significance
simple random sample, 41-2, 248
simulation, 216
skew distribution, 157, 160
social survey, see surveys
sorting variable, 112
sound effects, radio, 12
SPSS, 94-5, 303
square-root, 241, 245
squared, as in 'x squared', 245
squared paper for analysing results, 86-9, 97, 181
standard deviation, 155-6
    calculation of, 259-60
    of normal distribution, 278
standard error, 146, 280
    calculation of, 248
Statistical Package for the Social Sciences, see SPSS
statistical significance, 147-51, 157-62, 218, 281
    calculation of, 250-9, 260-71
statisticians, see experts on research
statistics
    books on, 298
    calculation of, 241-71
    concepts, 141-62
    official, 24, 44-6, 99-101, 216
    tables of, 302, 307-12
    theory, 272-82
status of correspondence students, 192-4
stem, in multiple-choice questions, 178
stock card, 294
stratified samples, 42-3, 44-6, 249-50
student advisers, 5, 13, 126, 188, 193
students taking correspondence courses
    difficulties, 9, 30, 102, 248-9
    examinations, 198-201
    experiment on, 126
    feedback from, 196-8
    profile, 188-91

  progress, 13, 21, 191-6
subjects, experimental, 129
subscript, 251
summation sign, 242, 245
summative evaluation, 227
  see also evaluation
surveys, social, 35-124
  administration, 72, 104
  basic principle, 35
  books on, 297
  examples, 8-15
  stages, 36, 123
  types, 21, 35-6
  uses, 3, 4
  see also analysis, checking, data-processing, fieldwork, questionnaires, reports, sampling
systematic random sample, 42, 248

t test
  for dependent sets of scores, 267
  for independent sets of scores, 261
  table of values, 309
tables
  layout of, 113-14, 142-5
  of expenditure, 288-92
  statistical, 302, 307-12
  use of in reports, 122, 180, 182
  see also crosstabs
tally-marks, 75-6, 88-9
tape for data storage, 94
tape-recording, 32-3
teacher training, evaluation of, 217-18
test of the difference between proportions, 250
tests
  in evaluation, 215
  in pre-testing, 175-80
  of statistical significance, see statistical significance
third world, see developing countries
three-way crosstab, 115
time-series, 209
time-sheet, 293
time taken by research, 25-6, 123, 172, 225-6
training courses
  evaluation of, 216
  for interviewers, 69-71
  preparatory research for, 9, 31
transcript of tape recording, 33
transfer cards, 77-86, 90, 97
translation, see language
tutors for correspondence courses, 192, 196-8
two-by-two table, 114, 115, 250, 255
two-choice questions, 177
two-way crosstab, 115

unimodal, 157, 275
unit costs, 295

value of a variable, 151, 242
variable, 112, 151, 242
variation, measures of, 154-6, 159
vegetable-growing booklet, 24, 212
visual aids, 9, 11, 27-8, 174, 180, 205

weak research designs in evaluation, 207, 211-12, 214-15
wheat campaign, 207-12
Wilcoxon matched-pairs signed-ranks test, 269-71, 311
word games, evaluation of, 222
writers of instructional materials, 5
    assessing written work, 163-71
    out of touch with readers, 28
    pre-testing material, 12, 172-3, 180, 234-5
    writing correspondence lessons, 10, 126-8, 131-2, 161

z, table of values for, 267